ARTIFICIAL INTELLIGENCE IN HEALTHCARE AND MEDICINE
ENHANCING THE EXPERT

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF ELECTRICAL ENGINEERING
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Andre Esteva
February 2018

This dissertation is online at: http://purl.stanford.edu/xg779kd4737

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Sebastian Thrun, Primary Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Stephen Boyd, Co-Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

John Duchi

Approved for the Stanford University Committee on Graduate Studies.

Patricia J. Gumport, Vice Provost for Graduate Education

This signature page was generated electronically upon submission of this dissertation in electronic format. An original signed hard copy of the signature page is on file in University Archives.

Preface

Due to a convergence of large open-source datasets, significant improvements in the parallelization capabilities of hardware (notably, multi-thousand core graphics processing units), and renewed academic interest in decades-old neural network algorithms, primary subfields of AI have flourished in the past 5 years. Natural Language Processing, Computer Vision, and robotics have attained impressive performance across many key AI tasks.

In this thesis we focus not on the development of new AI algorithms, but on their application to a series of important problems in healthcare and medicine. Machine learning and AI will impact industries that generate significant amounts of data, by extracting insights that humans cannot. Healthcare and medicine are examples of such industries and thus are excellent use cases for AI deployment.

Our story is divided into 4 sections - neuroscience, psychiatry, drug screening, and dermatology - all linked by the common thread of using AI to *enhance the expert*, either in-clinic or in the analysis of data. This underlying motif is the connection to a paradigm in AI development popularized as the *virtual cycle of AI*: build a product that has front-facing user value, generate data with it, use that data to train AI algorithms that improve your product, acquire more data, etc.

Acknowledgments

People often ask me what it's like to get a PhD. I like to draw an analogy to the outdoors. Getting a bachelors or masters degree is a bit like hiking a trail. First off, there is a trail in front of you that others have walked before. There is a start, a stop, and mileposts on the way. You have a map, and can track your progress over time and distance. Some trails are harder than others, but in essence they are the same. A PhD is like being dropped in the middle of the jungle, in a remote area all by yourself, and being asked to carve your way out. You must develop your own tools and skillsets to navigate the wilderness. You rarely know if you are headed in the right direction, and you often realize you're going in the wrong direction, or in circles. Given enough persistence, and improved skillsets over time, you eventually manage to make your way out of the jungle. One of the great challenges is the psychological pressure of not knowing when you will finish, if you will finish, or if you even want to finish. While this is fundamentally independent, it is nearly impossible to do without the support of those in your community. I could not have completed this journey without my inspirational advisor, my brilliant colleagues, the best friends and family in the world, and the most amazing parents anyone could ask for.

I would like to thank my advisor, Professor Sebastian Thrun. We met some years ago over a cup of coffee, discussed a number of exciting ideas across a range of fields, and it became clear that we both were interested in solving large-scale societal challenges using AI. It has been a pleasure and a privelege to be mentored by Sebastian. He has always encouraged me to focus on the grand vision of what we are trying to solve.

I would like to thank the numerous brilliant colleagues with whom I have had the pleasure of authoring papers and discussing ideas. My medical co-authors have ensured that the problems we tackle are relevant and mission-critical to the future of healthcare. My engineering co-authors have worked with me to build new technologies and push forth the state of the art.

I am indebted to the friends and family that were with me through this five-year journey. From late nights in the library and in lab to forcing me to get off campus and take breaks, to helping me remember just how important our work can be, I would not have gotten through this without their emotional support.

Finally, I owe the greatest debt of gratitude to my parents. They have given me everything, and everything I have achieved, I owe first and foremost to them.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

This thesis is composed of four chapters in neuroscience, psychiatry, drug screening, and dermatology - all linked by the common thread of using AI to *enhance the expert*.

In Chapter 2: Neuroscience, we show that humans categorize visual scenes based on the *functionality* of the scene, and that our brain's visual system is split into two networks: the first processes visual features, and the second connects information about the current scene with a broader temporal and spatial context. We go on to make a case for the scalability of neuroscience, adopting a paradigm focused on the recording of in-vivo information and the reporting of that information to external sources, contemplating how techniques will change as our ability to measure and interpret the activity of individual neurons grows to encompass the entire brain. We highlight a number of critical fields and technologies which we believe will be paramount to this topic.

In Chapter 3: Psychiatry, we demonstrate that an AI system composed of a camera and eye-tracker can classify Fragile-X-Syndrome Autism against generic developmental disorders directly from a video stream, by monitoring the interaction between a patient and physician. This system can diagnose autism in under 60 seconds, orders of magnitude faster than the hours to days of psychologist-time typically required.

Chapter 4: Drug Screening focuses on cellular biology, where we build a computer vision system capable of predicting molecular stain protocols of neurons in petri dishes, imaged via bright-field and confocal microscopy. This automation has the potential to significantly increase the iterative speed of drug trials. In experiments where neurons are perturbed with drugs (i.e. for Alzheimer's Disease) and stained - a process which disrupts the life-cycle of the neurons and can harm or kill them - we offer an equally accurate technique which passively monitors them and provides the same degree of insight. We name this technique *In-Silico Labeling*.

Finally, in Chapter 5: Dermatology, we apply computer vision to the classification, detection, and tracking of skin cancers in clinical and dermoscopic photographs. We begin by demonstrating that a deep neural network can achieve the performance of board-certified dermatologists at classifying skin cancers against benign neoplasms. We then build a system for the detection and tracking of single lesions, both cancerous

and benign, from far-away photographs of body sections.

1.0.1 Neuroscience

How do we know that a kitchen is a kitchen by looking? Traditional models posit that scene categorization is achieved through recognizing necessary and sufficient features and objects, yet there is little consensus about what these may be. However, scene categories should reflect how we use visual information. We therefore test the hypothesis that scene categories reflect functions, or the possibilities for actions within a scene. Our approach is to compare human categorization patterns with predictions made by both functions and alternative models. We collected a large-scale scene category distance matrix (5 million trials) by asking observers to simply decide whether two images were from the same or different categories. Using the actions from the American Time Use Survey, we mapped actions onto each scene (1.4 million trials). We found a strong relationship between ranked category distance and functional distance (r=0.50, or 66% of the maximum possible correlation). The function model outperformed alternative models of object-based distance (r=0.33), visual features from a convolutional neural network (r=0.39), lexical distance (r=0.27), and models of visual features. Using hierarchical linear regression, we found that functions captured 85.5% of overall explained variance, with nearly half of the explained variance captured only by functions, implying that the predictive power of alternative models was due to their shared variance with the function-based model. These results challenge the dominant school of thought that visual features and objects are sufficient for scene categorization, suggesting instead that a scenes category may be determined by the scenes function.

Research on visual scene understanding has identified a number of regions involved in processing natural scenes, but has lacked a unifying framework for understanding how these different regions are organized and interact. We propose a new organizational principle, in which scene processing relies on two distinct networks that split the classically defined Parahippocampal Place Area (PPA). The first network consists of the Transverse Occipital Sulcus (TOS, or the Occipital Place Area) and the posterior portion of the PPA (pPPA). These regions have a well-defined retinotopic organization and do not show strong memory or context effects, suggesting that this network primarily processes visual features from the current view of a scene. The second network consists of the caudal Inferior Parietal Lobule (cIPL), Retrosplenial Cortex (RSC), and the anterior portion of the PPA (aPPA). These regions are involved in a wide range of both visual and non-visual tasks involving episodic memory, navigation, and imagination, and connect information about a current scene view with a much broader temporal and spatial context. We provide evidence for this division from a diverse set of sources. Using a data-driven approach to parcellate resting-state fMRI data, we identify coherent functional regions corresponding to scene-processing areas. We then show that a network clustering analysis separates these scene-related regions into two adjacent networks, which show sharp changes in connectivity properties. Additionally, we argue that the cIPL has been previously overlooked as a critical region for full scene understanding, based on a meta-analysis of previous functional studies as well as diffusion tractography results showing that cIPL is well-positioned to connect visual cortex with other cortical systems. This new framework for understanding the neural substrates of scene processing bridges results from many lines of

research, and makes specific predictions about functional properties of these regions. Two major initiatives to accelerate research in the brain sciences have focused attention on developing a new generation of scientific instruments for neuroscience. These instruments will be used to record static (structural) and dynamic (behavioral) information at unprecedented spatial and temporal resolution and report out that information in a form suitable for computational analysis. We distinguish between recording — taking measurements of individual cells and the extracellular matrix — and reporting — transcoding, packaging and transmitting the resulting information for subsequent analysis — as these represent very different challenges as we scale the relevant technologies to support simultaneously tracking the many neurons that comprise neural circuits of interest. We investigate a diverse set of technologies with the purpose of anticipating their development over the span of the next 10 years and categorizing their impact in terms of short-term [1-2 years], medium-term [2-5 years] and longer-term [5-10 years] deliverables.

This chapter is joint work with Chris Baldassano, Michelle Greene, Diane M. Beck, Fei-Fei Li, Thomas Dean, Biafra Ahanonu, Mainak Chowdhury, Anjali Datta, Daniel Eth, Nobie Redmon, Oleg Rumyantsev, and Ysis Tarter.

1.0.2 Psychiatry

This section proposes a system for fine-grained classification of developmental disorders via measurements of individuals eye-movements using multi-modal visual data. While the system is engineered to solve a psychiatric problem, we believe the underlying principles and general methodology will be of interest not only to psychiatrists but to researchers and engineers in medical machine vision. The idea is to build features from different visual sources that capture information not contained in either modality. Using an eye-tracker and a camera in a setup involving two individuals speaking, we build temporal attention features that describe the semantic location that one person is focused on relative to the other persons face. In our clinical context, these temporal attention features describe a patients gaze on finely discretized regions of an interviewing clinicians face, and are used to classify their particular developmental disorder.

This chapter is joint work with Guido Pusiol, Fei-Fei Li, Arnold Milstein, Mike Frank, and Scott Hall.

1.0.3 Drug-Screening

Imaging is a central method in life sciences, and the drive to extract information from microscopy approaches has led to methods to fluorescently label specific cellular constituents. However, the specificity of fluorescent labels varies, labeling can confound biological measurements, and spectral overlap limits the number of labels to a few that can be resolved simultaneously. Here, we developed a deep learning computational approach called in silico labeling" (ISL) that reliably infers information from unlabeled fixed or live biological samples that would normally require invasive labeling. ISL predicts different labels in multiple cell types from independent laboratories. It makes cell type and state predictions by integrating in silico labels, and is not limited by spectral overlap. The network exhibits transfer learning, enabling it to adapt to new samples with

small training datasets. ISL thus provides, for negligible additional cost, biological insights from images of unlabeled samples that would be undesirable or impossible to acquire otherwise.

This chapter is joint work with Eric Christiansen, Samuel J. Yang, D. Michael Ando, Ashkan Javaherian, Gaia Skibinski, Scott Lipnick, Elliot Mount, Alison ONeil, Kevan Shah, Alicia K. Lee, Piyush Goyal, Liam Fedus, Ryan Poplin, Marc Berndl, Lee L. Rubin, Philip Nelson, and Steven Finkbeiner.

1.0.4 Dermatology

Skin cancer - the most common human malignancy [1] [352] [401] - is primarily diagnosed visually, beginning with an initial clinical screening, followed potentially by dermoscopic analysis, a biopsy, and histopathological examination. Automated classification of skin lesions using images is a challenging task due to the fine-grained variability of skin lesion appearance. Deep convolutional neural networks (CNN) [246] [248] show great promise for general and highly variable tasks over many fine-grained object categories [361] [233] [189] [410] [411] [171]. Here we show classification of skin lesions using a single CNN, trained end-to-end directly from images using only their pixels and disease labels as inputs. We train a CNN on a dataset of 129,450 clinical images - two orders of magnitude larger than previous datasets [275] - consisting of 2,032 different diseases. We test its performance against 21 board-certified dermatologists on biopsy-proven clinical images with two critical use cases: binary classification of (1) malignant carcinomas versus benign seborrheic keratoses, and (2) malignant melanomas versus benign nevi. Case (1) represents the identification of the most common cancers, and case (2) represents identification of the deadliest skin cancer. The CNN achieves performance on par with all tested experts across both tasks, demonstrating, for the first time, an artificial intelligence with dermatologist-level skin cancer classification capability. It is projected that 6.3 billion smartphone subscriptions will exist by the year 2021 [77]. Outfitted with deep neural networks, mobile devices can extend the reach of dermatologists outside of the clinic, and enable low-cost universal access to vital diagnostic care.

Dense object detection and temporal tracking is needed across applications domains ranging from people-tracking to analysis of satellite imagery over time. The detection and tracking of malignant skin cancers and benign moles poses a particularly challenging problem due to the general uniformity of large skin patches, the fact that skin lesions vary little in their appearance, and the relatively small amount of data available. Here we introduce a novel data synthesis technique that merges images of individual skin lesions with full-body images and heavily augments them to generate significant amounts of data. We build a convolutional neural network (CNN) based system, trained on this synthetic data, and demonstrate superior performance to traditional detection and tracking techniques. Additionally, we compare our system to humans trained with simple criteria. Our system is intended for potential clinical use to augment the capabilities of healthcare providers. While domain-specific, we believe the methods invoked in this work will be useful in applying CNNs across domains that suffer from limited data availability.

This chapter is joint work with Sebastian Thrun, Rob Novoa, Justin Ko, Yunzhu Li, Brett Kuprel, Helen Blau, and Susan Swetter.

Chapter 2

Neuroscience

2.1 Visual Scenes are Categorized by Function

Although more than half a century has passed since Attneave issued this challenge, we still have little understanding of how we categorize and conceptualize visual content. The notion of similarity, or family resemblance, is implicit in how content is conceptualized (Wittgenstein, 2010), yet similarity cannot be defined except in reference to a feature space to be operated over [147, 278]. What feature spaces determine environmental categories? Traditionally, it has been assumed that this feature space is comprised of a scenes component visual features and objects [36, 56, 273, 348, 397]. Mounting behavioral evidence, however, indicates that human observers have high sensitivity to the global meaning of an image [128, 152, 332], and very little sensitivity to the local objects and features that are outside the focus of attention [346]. Consider the image of the kitchen in Figure 2.1. If objects determine scene category membership, then we would expect the kitchen supply store (left) to be conceptually equivalent to the kitchen. Alternatively, if scenes are categorized (labeled) according to spatial layout and surfaces [24, 314, 420], then observers might place the laundry room (center) into the same category as the kitchen. However, most of us share the intuition that the medieval kitchen (right) is in the same category, despite sharing few objects and features with the top image. Why is the image on the right a better category match to the modern kitchen than the other two?

Figure 2.8 illustrates our approach. We constructed a large-scale scene category distance matrix by querying over 2,000 observers on over 63,000 images from 1055 scene categories (Figure 2.8A). Here, the distance between two scene categories was proportional to the number of observers who indicated that the two putative categories were different. We compared this human categorization pattern with an function-based pattern created by asking hundreds of observers to indicate which of several hundred actions could take place in each scene (Figure 2.8B). We can then compute the function-based distance for each pair of categories. We found a striking resemblance between function-based distance and the category distance pattern. The function model not only explained more variance in the category distance matrix than leading models of visual features and objects, but also contributed the most uniquely explained variance of any tested model. These results suggest

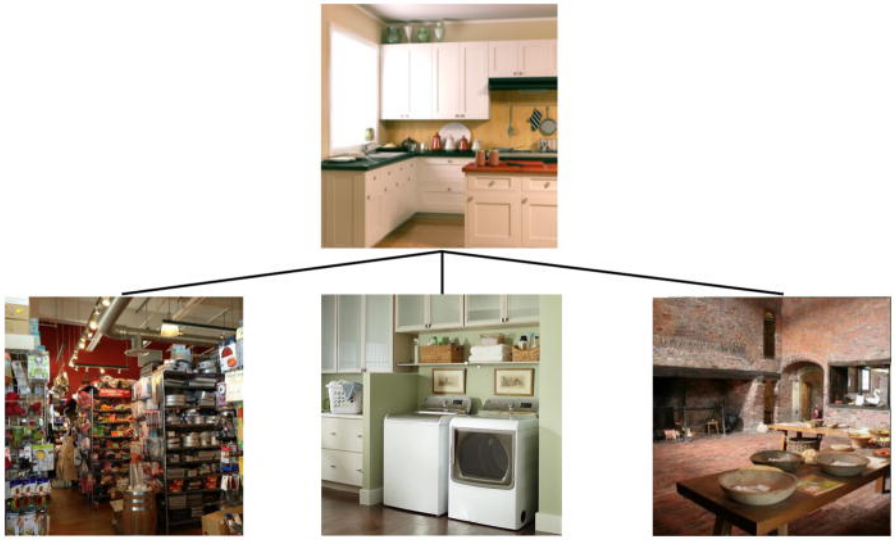


Figure 2.1: The top image depicts a kitchen. Which of the bottom images is also a kitchen? Many influential models of visual categorization assume that scenes sharing objects, such as the kitchen supply store (left), or layout, such as the laundry room (middle) would be placed into the same category by human observers. Why is the medieval kitchen also a kitchen despite having very different objects and features from the top kitchen?

that a scenes functions provide a fundamental coding scheme for human scene categorization. In other words, of the models tested, the functions afforded by the scene best explains why we consider two images to be from the same category.

2.1.1 Methods

Creating Human Scene Category Distance Matrix

The English language has terms for hundreds of types of environments, a fact reflected in the richness of large-scale image databases such as ImageNet [102] or SUN [460]. These databases used the WordNet [285] hierarchy to identify potential scene categories. Yet we do not know how many of these categories reflect basic- or entry-level scene categories, as little is known about the hierarchical category structure of scenes [425]. Therefore, our aim was to discover this category structure for human observers at a large scale.

To derive a comprehensive list of scene categories, we began with a literature review. Using Google Scholar, we identified 116 papers in human visual cognition, cognitive neuroscience, or computer vision matching the keywords scene categorization or scene classification that had a published list of scene categories. 1535 unique category terms were identified over all papers. Our goal was to identify scene categories with at least 20 images in publically available databases. We removed 204 categories that did not meet this criterion. We then removed categories describing animate entities (e.g. Crowd of people, N=44); specific places (e.g. Alaska, N=42); events (e.g. forest fire, N=35); or objects (e.g. playing cards, N=93). Finally,

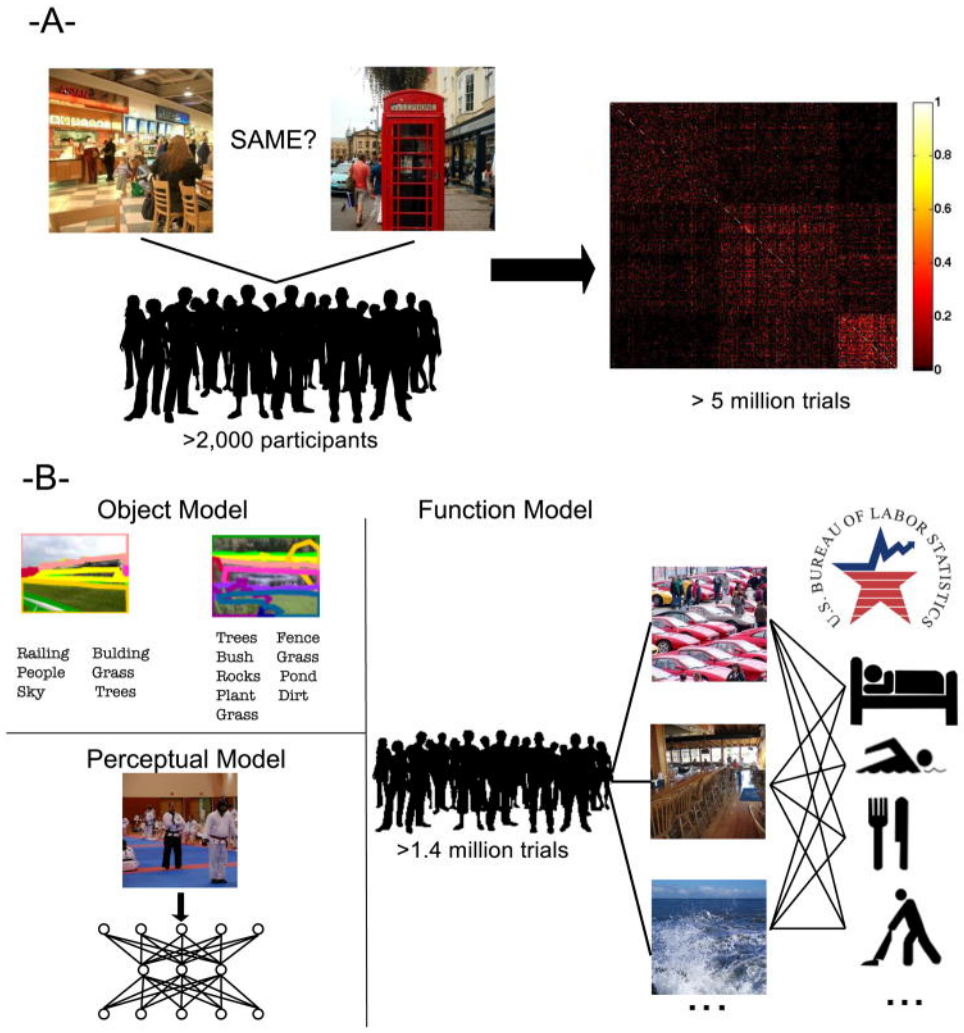


Figure 2.2: (A) We used a large-scale online experiment to generate a distance matrix of scene categories. Over 2,000 individuals viewed more than 5 million trials in which participants viewed two images and indicated whether they would place the images into the same category. (B) Using the LabelMe tool [362] we examined the extent to which scene category similarity was related to scenes having similar objects. Our perceptual model used the output features of a state-of-the-art convolutional neural network [383], to examine the extent to which visual features contribute to scene category. To generate the functional model, we took 227 actions from the American Time Use Survey. Using crowdsourcing, participants indicated which actions could be performed in which scene categories.

we omitted 62 categories for being close synonyms of another (e.g. country and countryside). This left us with a total of 1055 scene categories. To obtain images for each category, 722 categories were found in the SUN database [459], 306 were taken from ImageNet [102], 24 from the Corel database, and three from the 15-scene database of [130, 242, 314].

We will refer to the 1,055 scene categories as putative categories. Good categories have both high within-category similarity (cohesion), as well as high between-category distance (distinctiveness) [190, 357]. We performed a large-scale experiment with over 2,000 human observers using Amazons Mechanical Turk (AMT). In each trial, two images were presented to observers side by side. Half of the image pairs came from the same putative scene category, while the other half were from two different categories that were randomly selected. Image exemplars were randomly selected within a category on each trial. In order to encourage participants to categorize at the basic- or entry-level [199, 425], we gave participants the following instructions: Consider the two pictures below, and the names of the places they depict. Names should describe the type of place, rather than a specific place and should make sense in finishing the following sentence I am going to the . , following the operational definition applied in the creation of the SUN database [459]. To ensure that the instructions were understood and followed, participants were also asked to type in the category name that they would use for the image on the left-hand side. These data were not analyzed. Participants were not placed under time pressure to respond, and images remained on screen until response was recorded.

Potential participants were recruited from a pool of trusted observers with at least 2,000 previously approved trials with at least 98% approval. Additionally, participants were required to pass a brief scene vocabulary test before participating. In the vocabulary test, potential participants were required to match ten scene images to their appropriate category name (see Supplementary Material for names and images). 245 potential participants attempted the qualification test and did not pass. Trials from 14 participants were omitted from analysis for inappropriate typing in the response box. Trials were omitted when workers pasted the image URL into the category box instead of providing a name (N=586 trials from 3 workers), for submitting the hit before all trials were complete (N=559 trials from 4 workers), for typing category names in languages other than English (N=195 trials from 2 workers), typing random character strings (N=111 trials from 2 workers), or for typing in words such as same, left, or pictures, implying that the instructions were not understood (N=41 trials from 3 workers). Workers were compensated $0.02 for each trial. We obtained at least 10 independent observations for each cell in the 1055 by 1055 scene matrix, for a total of over 5 million trials. Individual participants completed a median of 5 hits of this task (range: 136,497). There was a median of 1,116 trials in each of the diagonal entries of the matrix, and a median of 11 trials in each cell of the off-diagonal entries.

From the distribution of same and different responses, we created a dissimilarity matrix in which the distance between two scene categories was defined as the proportion of participants who indicated that the two categories were different. From the 1,055 categories, we identified 311 categories with the strongest within-category cohesion (at least 70% of observers agreed that images were from the same category). In general, categories that were omitted were visually heterogeneous, such as community center, or were inherently multimodal. For example, dressing room could reflect the backstage area of a theatre, or a place to try on clothes in a department store. Thus, the final dataset included 311 scene categories from 885,968 total trials, and from 2,296 individual workers.

Creating the Scene Function Spaces

In order to determine whether scene categories are governed by functions, we needed a broad space of possible actions that could take place in our comprehensive set of scene categories. We gathered these actions from the lexicon of the American Time Use Survey (ATUS), a project sponsored by the US Bureau of Labor Statistics that uses U.S. census data to determine how people distribute their time across a number of activities. The lexicon used in this study was pilot tested over the course of three years (Shelley, 2005), and therefore represents a complete set of goal-directed actions that people can engage in. This lexicon was created independently from any question surrounding vision, scenes, or categories, therefore avoiding the potential problem of having functions that were designed to distinguish among categories of visual scenes. Instead, they simply describe common actions one can engage in in everyday life. The ATUS lexicon includes 428 specific activities organized into 17 major activity categories and 105 mid-level categories. The 227 actions included in our study included the most specific category levels with the following exceptions:

The superordinate category Caring for and Helping Non-household members was dropped as these actions would be visually identical to those in the Caring for and Helping Household members category. In the ATUS lexicon, the superordinate-level category Work contained only two specific categories (primary and secondary jobs). Because different types of work can look very visually different, we expanded this category by adding 22 categories representing the major labor sectors from the Bureau of Labor Statistics. The superordinate-level category Telephone calls was collapsed into one action because we reasoned that all telephone calls would look visually similar. The superordinate-level category Traveling was similarly collapsed into one category because being in transit to go to school (for example) should be visually indistinguishable from being in transit to go to the doctor. All instances of Security procedures have been unified under one category for similar reasons. All instances of Waiting have been unified under one category. All Not otherwise specified categories have been removed. The final list of actions can be found in the Supplemental Materials.

To compare this set of comprehensive functions to a human-generated list of functions applied to visual scenes, we took the 36 function/affordance rankings from the SUN attribute database [324]. In this set, observers were asked to generate attributes that differentiated scenes.

Mapping Functions Onto Images

In order to test our hypothesis that scene category distance is reflected in the distance of scenes functions, we need to map functions onto scene categories. Using a separate large-scale online experiment, 484 participants indicated which of the 227 actions could take place in each of the 311 scene categories. Participants were screened using the same criterion described above. In each trial, a participant saw a randomly selected exemplar image of one scene category along with a random selection of 17 or 18 of the 227 actions. Each action was hyperlinked to its description in the ATUS lexicon. Participants were instructed to use check boxes to indicate which of the actions would typically be done in the type of scene shown.

Each individual participant performed a median of 9 trials (range: 14,868). Each scene category function pair was rated by a median of 16 participants (range: 486), for a total of 1.4 million trials.

We created a 311-category by 227-function matrix in which each cell represents the proportion of participants indicating that the action could take place in the scene category. Since scene categories varied widely in the number of actions they afford, we created a distance matrix by computing the cosine distance between all possible pairs of categories, resulting in a 311311 function-based distance matrix. This measures the overlap between actions while being invariant to the absolute magnitude of the action vector.

Function Space MDS Analysis

To better understand the scene function space, we performed a classical metric multidimensional scaling (MDS) decomposition of the function distance matrix. This yielded an embedding of the scene categories such that inner products in this embedding space approximate the (double-centered) distances between scene categories, with the embedding dimensions ranked in order of importance [55]. In order to better understand the MDS dimensions, we computed the correlation coefficient between each action (across scene categories) with the category coordinates for a given dimension. This provides us with the functions that are the most and least associated with each dimension.

Alternative Models

To put the performance of the function-based model in perspective, we compared it to nine alternative models based on previously proposed scene category primitives. Five of the models represented visual features, one model considered human-generated scene attributes, and one model examined the human-labeled objects in the scenes. As with the function model, these models yielded scene category by feature matrices that were converted to distance matrices using cosine distance, and then compared to the category distance matrix. The object and attribute models, like the functional model, were created from human observers scene labeling. Additionally, two models measured distances directly, based either on the lexical distance between scene category names (the Semantic Model), or simply by whether scenes belonged to the same superordinate level category (indoor, urban or natural; the Superordinate-Category Model). We will detail each of the models below.

Models of Visual Features

A common framework for visual categorization and classification involves finding the necessary and sufficient visual features to perform categorization e.g. [131, 243, 315, 345, 440]. Here we constructed distance matrices based on various visual feature models to determine how well they map on the human categorization (i.e. the category dissimilarity matrix) and in particular compare their performance to our functional category model.

Convolutional Neural Network

In order to represent the state-of-the-art in terms of visual features, we generated a visual feature vector using the publicly distributed OverFeat convolutional neural network (CNN) [382], which was trained on the ImageNet 2012 training set [102]. These features, computed by iteratively applying learned nonlinear filters to the image, have been shown to be a powerful image representation for a wide variety of visual tasks [342]. This 7-layer CNN takes an image of size 231231 as input, and produces a vector of 4096 image features that are optimized for 1000-way object classification. This network achieves top-5 object recognition on ImageNet 2012 with approximately 16% error, meaning that the correct object is one of the models first five responses in 84% of trials. Using the top layer of features, we averaged the features for all images in each scene category to create a 311-category by 4096-feature matrix.

Gist

We used the Gist descriptor features of [315]. This popular model for scene recognition provides a summary statistic representation of the dominant orientations and spatial frequencies at multiple scales coarsely localized on the image plane. We used spatial bins at 4 cycles per image and 8 orientations at each of 4 spatial scales for a total of 3,072 filter outputs per image. We averaged the gist descriptors for each image in each of the 311 categories to come up with a single 3,072-dimensional descriptor per category.

Color histograms

In order to determine the role of color similarity in scene categorization, we represented color using LAB color space. For each image, we created a two-dimensional histogram of the a* and b* channels using 50 bins per channel. We then averaged these histograms over each exemplar in each category, such that each category was represented as a 2500 length vector representing the averaged colors for images in that category. The number of bins was chosen to be similar to those used in previous scene perception literature [313].

Tiny Images

Torralba and colleagues [420] demonstrated that human scene perception is robust to aggressive image downsampling, and that an image descriptor representing pixel values from such downsampled images could yield good results in scene classification. Here, we downsampled each image to 32 by 32 pixels (grayscale). We created our 311-category by 1024 feature matrix by averaging the downsampled exemplars of each category together.

Gabor Wavelet Pyramid

To assess a biologically inspired model of early visual processing, we represented each image in this database as the output of a bank of multi-scale Gabor filters. This type of representation has been used to successfully model the representation in early visual areas [210]. Each image was converted to grayscale, down sampled

to 128 by 128 pixels, and represented with a bank of Gabor filters at three spatial scales (3, 6 and 11 cycles per image with a luminance-only wavelet that covers the entire image), four orientations (0, 45, 90 and 135 degrees) and two quadrature phases (0 and 90 degrees). An isotropic Gaussian mask was used for each wavelet, with its size relative to spatial frequency such that each wavelet has a spatial frequency bandwidth of 1 octave and an orientation bandwidth of 41 degrees. Wavelets were truncated to lie within the borders of the image. Thus, each image is represented by 3*3*2*4+6*6*2*4+11*11*2*4 = 1328 total Gabor wavelets. We created the feature matrix by averaging the Gabor weights over each exemplar in each category.

Object-based Model

Our understanding of high-level visual processing has generally focused on object recognition, with scenes considered as a structured set of objects [36]. Therefore, we also consider a model of scene categorization that is explicitly built upon objects. In order to model the similarity of objects within scene categories, we employed the LabelMe tool [363] that allows users to outline and annotate each object in each image by hand. 7,710 scenes from our categories were already labeled in the SUN 2012 release [459], and we augmented this set by labeling an additional 223 images. There were a total of 3,563 unique objects in this set. Our feature matrix consisted of the proportion of scene images in each category containing a particular object. For example, if 10 out of 100 kitchen scenes contained a blender, the entry for kitchen-blender would be 0.10. In order to estimate how many labeled images we would need to robustly represent a scene category, we performed a bootstrap analysis in which we resampled the images in each category with replacement (giving the same number of images per category as in the original analysis), and then measured the variance in distance between categories. With the addition of our extra images, we ensured that all image categories either had at least 10 fully labeled images or had mean standard deviation in distance to all other categories of less than 0.05 (e.g. less than 5% of the maximal distance value of 1).

Scene-Attribute Model

Scene categories from the SUN database can be accurately classified according to human-generated attributes that describe a scenes material, surface, spatial, and functional scene properties [324]. In order to compare our function-based model to another model of human-generated attributes, we used the 66 non-function attributes from [324] for the 297 categories that were common to our studies. To further test the role of functions, we then created a separate model from the 36 function-based attributes from their study. These attributes are listed in the Supplementary Material.

Semantic Models

Although models of visual categorization tend to focus on the necessary features and objects, it has long been known that most concepts cannot be adequately expressed in such terms [456]. As semantic similarity has been suggested as a means of solving category induction [240], we examined the extent to which category

structure follows from the semantic similarity between category names. We examined semantic similarity by examining the shortest path between category names in the WordNet tree using the Wordnet::Similarity implementation of [325]. The similarity matrix was normalized and converted into distance. We examined each of the metrics of semantic relatedness implemented in Wordnet::Similarity and found that this path measure was the best correlated with human performance.

Superordinate-Category Model

As a baseline model, we examined how well a model that groups scenes only according to superordinate-level category would predict human scene category assessment. We assigned each of the 311 scene categories to one of three groups (natural outdoors, urban outdoors or indoor scenes). These three groups have been generally accepted as mutually exclusive and unambiguous superordinate-level categories [425, 459]. Then, each pair of scene categories in the same group was given a distance of 0 while pairs of categories in different groups were given a distance of 1.

Model Assessment

To assess how each of the feature spaces resembles the human categorization pattern, we created a 311311 distance matrix representing the distance between each pair of scene categories for each feature space. We then correlated the off-diagonal entries in this distance matrix with those of the category distance matrix from the scene categorization experiment. Since these matrices are symmetric, the off-diagonals were represented in a vector of 48,205 distances.

Noise Ceiling

The variability of human categorization responses puts a limit on the maximum correlation expected by any of the tested models. In order to get an estimate of this maximum correlation, we used a bootstrap analysis in which we sampled with replacement observations from our scene categorization dataset to create two new datasets of the same size as our original dataset. We then correlated these two datasets to one another, and repeated this process 1000 times.

Hierarchical Regression Analysis

In order to understand the unique variance contributed by each of our feature spaces, we used hierarchical linear regression analysis, using each of the feature spaces both alone and in combination to predict the human categorization pattern. In total, 15 regression models were used: (1) all feature spaces used together; (2) the top four performing features together (functions, objects, attributes and the CNN visual features); (36) each of the top four features alone; (611) each pair of the top four features; (1215) each set of three of the top four models. By comparing the r2 values of a feature space used alone to the r2 values of that space in conjunction with another feature space, we can infer the amount of variance that is independently explained

by that feature space. In order to visualize this information in an Euler diagram, we used EulerAPE software [282].

2.1.2 Results

Human Scene Category Distance

To assess the conceptual structure of scene environments, we asked over 2,000 human observers to categorize images as belonging to 311 scene categories in a large-scale online experiment. The resulting 311 by 311 category distance matrix is shown in Figure 2.3. In order to better visualize the category structure, we have ordered the scenes using the optimal leaf ordering for hierarchical clustering [25]; allowing us to see what data-driven clusters emerge.

Several category clusters are visible. Some clusters appear to group several subordinate-level categories into a single entry-level concept, such as bamboo forest, woodland and rainforest being examples of forests. Other clusters seem to reflect broad classes of activities (such as sports) which are visually heterogeneous and cross other previously defined scene boundaries, such as indoor-outdoor [129, 177, 415, 425], or the size of the space [151, 315, 320]. Such activity-oriented clusters hint that the actions that one can perform in a scene (the scenes functions) could provide a fundamental grouping principle for scene category structure.

Function-based Distance Best Correlates with Human Category Distance

For each of our feature spaces, we created a distance vector (see Model Assessment) representing the distance between each pair of scene categories. We then correlated this distance vector with the human distance vector from the previously described experiment.

In order to quantify the performance of each of our models, we defined a noise ceiling based on the inter-observer reliability in the human scene distance matrix. This provides an estimate of the explainable variance in the scene categorization data, and thus provides an upper bound on the performance of any of our models. Using bootstrap sampling (see Methods), we found an inter-observer correlation of r=0.76. In other words, we cannot expect a correlation with any model to exceed this value.

Function-based similarity had the highest resemblance to the human similarity pattern (r=0.50 for comprehensive set, and r=0.51 for the 36 functional attributes). This represents about 2/3 of the maximum observable correlation obtained from the noise ceiling. As shown in Figure 2.4, this correlation is substantially higher than any of the alternative models we tested. The two function spaces were highly correlated with one another (r=0.63). As they largely make the same predictions, we will use the results from the 227-function set for the remainder of the paper.

Of course, being able to perform similar actions often means manipulating similar objects, and scenes with similar objects are likely to share visual features. Therefore, we compared function-based categorization patterns to alternative models based on perceptual features, non-function attributes, object-based similarity, and the lexical similarity of category names.

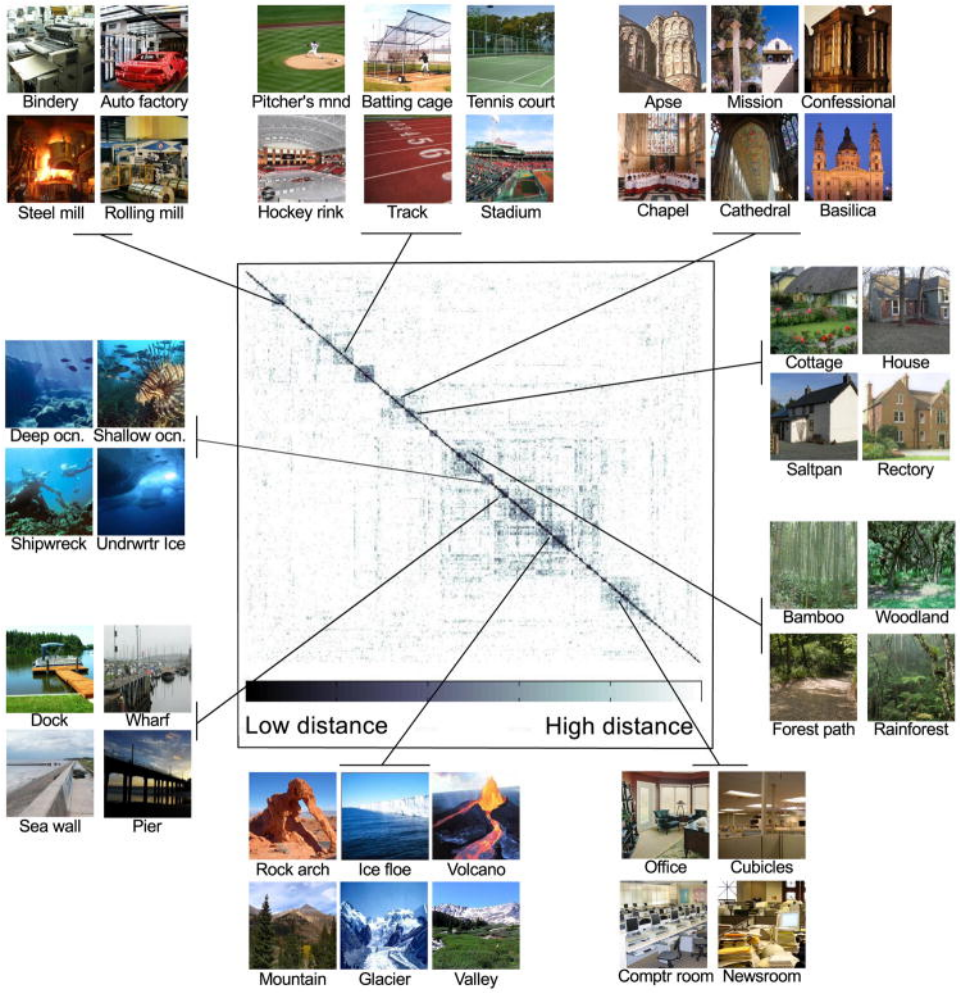


Figure 2.3: The human category distance matrix from our large-scale online experiment was found to be sparse. Over 2,000 individual observers categorized images in 311 scene categories. We visualized the structure of this data using optimal leaf ordering for hierarchical clustering, and show representative images from categories in each cluster.

We tested five different models based on purely visual features. The most sophisticated used the top-level features of a state-of-the-art convolutional neural network model (CNN), [382] trained on the ImageNet database [102]. Category distances in CNN space produced a correlation with human category dissimilarity of r=0.39. Simpler visual features, however, such as gist [315], color histograms [313], Tiny Images [421], and wavelets [210] had low correlations with human scene category dissimilarity.

Category structure could also be predicted to some extent based on the similarity between the objects present in scene images (r=0.33, using human-labeled objects from the LabelMe database, [363], the non

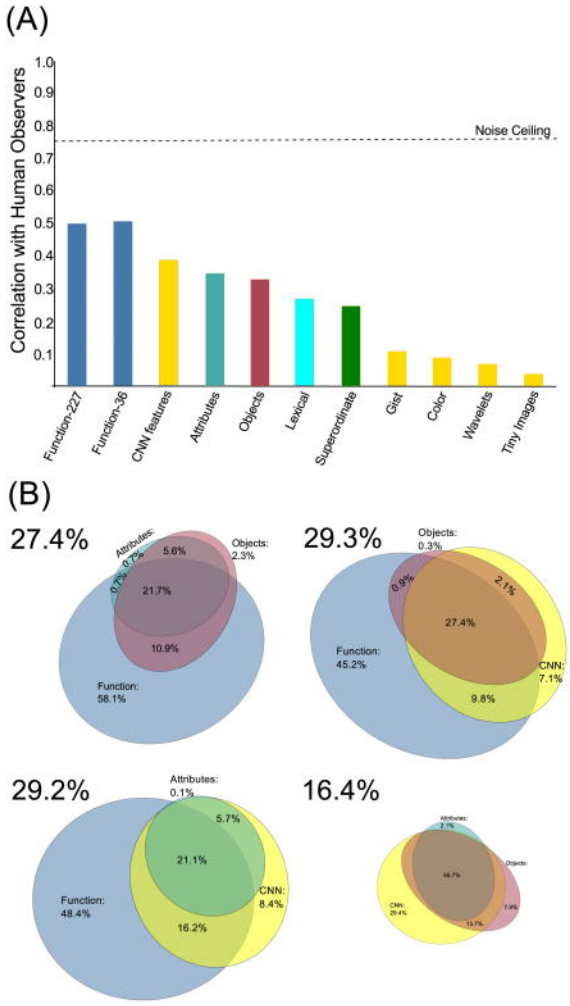


Figure 2.4: (A) Correlation of all models with human scene categorization pattern. Function-based models (dark blue, left) showed the highest resemblance to human behavior, achieving 2/3 of the maximum explainable similarity (black dotted line). Of the models based on visual features (yellow), only the model using the top-level features of the convolutional neural network (CNN) showed substantial resemblance to human data. The object-based model, the attribute-based model, the lexical model and the superordinate-level model all showed moderate correlations. (B) Euler diagrams showing the distribution of explained variance for sets of the four top-performing models. The function-based model (comprehensive) accounted for between 83.3% and 91.4% of total explained variance of joint models, and between 45.2% and 58.1% of this variance was not shared with alternative models. Size of Euler diagrams is approximately proportional to the total variance explained.

function-based attributes (r=0.28) of the SUN attribute database [324], or the lexical distance between category names in the WordNet tree [187, 285, 325] (r=0.27). Surprisingly, a model that merely groups scenes by superordinate-level categories (indoor, urban or natural environments) also had a sizeable correlation (r=0.25)

with human dissimilarity patterns.

Although each of these feature spaces had differing dimensionalities, this pattern of results also holds if the number of dimensions is equalized through principal components analysis. We created minimal feature matrices by using the first N PCA components, and then correlated the cosine distance in these minimal feature spaces with the human scene distances, see Figure 2.5. We found that the functional features were still the most correlated with human behavior.

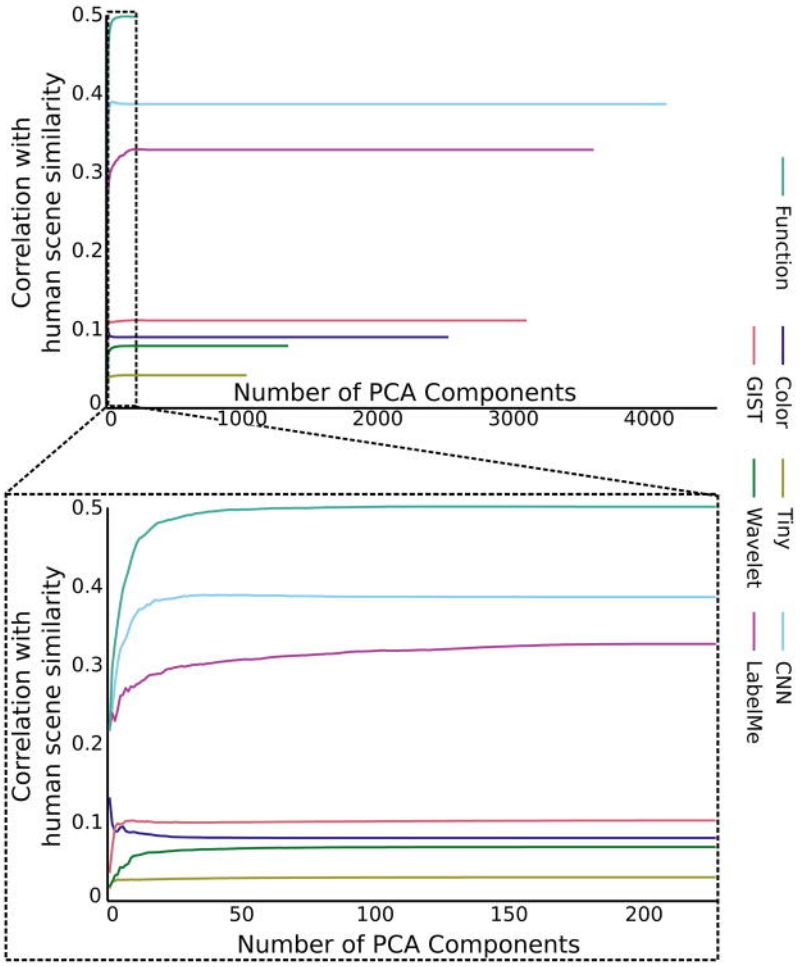


Figure 2.5: Robustness to dimensionality reduction. For each feature space, we reconstructed the feature matrix using a variable number of PCA components and then correlated the cosine distance in this feature space with the human scene distances. Although the number of features varies widely between spaces, all can be described in 100 dimensions, and the ordering of how well the features predict human responses is essentially the same regardless of the number of original dimensions.

Table 2.1: Variance explained (r^2) by fifteen regression models

Model	r^2
Attribute	0.08
Object	0.11
CNN	0.15
Function	0.25
Object + Attribute	0.11
Attribute + CNN	0.15
Object + CNN	0.16
Object + Function	0.27
Attribute + Function	0.27
CNN + Function	0.29
Object + Attribute + CNN	0.16
Object + Attribute + Function	0.27
Attribute + CNN + Function	0.29
Object + CNN + Function	0.29
Attribute + Object + CNN + Function	0.29

Independent Contributions from Alternative Models

To what extent does function-based similarity uniquely explain the patterns of human scene categorization? Although function-based similarity was the best explanation of the human categorization pattern of all the models we tested, CNN and object-based models also had sizeable correlations with human behavior. To what extent do these models make the same predictions?

In order to assess the independent contributions made by each of the models, we used a hierarchical linear regression analysis in which each of the three top-performing models was used either separately or in combination to predict the human similarity pattern. By comparing the r2 values from the individual models to the r2 values for the combined model, we can assess the unique variance explained by each descriptor. A combined model with all features explained 31% of the variance in the human similarity pattern (r=0.56). This model is driven almost entirely by the top four feature spaces (functions, CNN, attribute, and object labels), which explained 95% of the variance from all features, a combined 29.4% of the total variance (r=0.54). Note that functions explained 85.6% of this explained variance, indicating that the object and perceptual features only added a small amount of independent information (14.4% of the combined variance). Variance explained by all 15 regression models is listed in Table 2.1.

Although there was a sizable overlap between the portions of the variance explained by each of the models (see Figure 2.4B), around half of the total variance explained can be attributed only to functions (44.2% of the explained variance in top four models), and was not shared by the other three models. In contrast, the independent variance explained by CNN features, object-based features, and attributes accounted for only 6.8%, 0.6%, and 0.4% of the explained variance respectively. Therefore, the contributions of visual, attribute, and object-based features are largely shared with function-based features, further highlighting the utility of functions for explaining human scene categorization patterns.

Table 2.2: Correlation of top-four models in each of the three superordinate-level scene categories. The function-based model performs similarly in all types of scenes, while the CNN, attribute, and object-based models perform poorly in indoor environments.

	Indoor	**Urban**	**Natural**
Functions	0.50	0.47	0.51
CNN	0.37	0.43	0.59
Attributes	0.15	0.20	0.41
Objects	0.19	0.27	0.44

Functions Explain All Types of Scene Categories

Does the impressive performance of the functional model hold over all types of scene categories, or is performance driven by outstanding performance on a particular type of scene? To address this question, we examined the predictions made by the three top-performing models (functions, CNN and objects) on each of the superordinate-level scene categories (indoor, urban and natural landscape) separately. As shown in Table 2.2, we found that the function-based model correlated similarly with human categorization in all types of scenes. This is in stark contrast to the CNN and object models, whose performance was driven by performance on the natural landscape scenes.

Examining Scene Function Space

In order to better understand the function space, we performed classical multi-dimensional scaling on the function distance matrix, allowing us to identify how patterns of functions contribute to the overall similarity pattern. We found that at least 10 MDS dimensions were necessary to explain 95% of the variance in the function distance matrix, suggesting that the efficacy of the function-based model was driven by a number of distinct function dimensions, rather than just a few useful functions. We examined the projection of categories onto the first three MDS dimensions. As shown in Figure 2.6, the first dimension appears to separate indoor locations that have a high potential for social interactions (such as socializing and attending meetings for personal interest) from outdoor spaces that afford more solitary activities, such as hiking and science work. The second dimension separates work-related activities from leisure. Later dimensions appear to separate environments related to transportation and industrial workspaces from restaurants, farming, and other food-related environments, see Figure 2.7 for listing of associated categories and functions for each MDS dimension. A follow-up experiment demonstrated that functions that are highly associated with a particular object (e.g. mailing is strongly associated with objects such as mailboxes and envelopes) are equally predictive of categorization patterns as functions that do not have strong object associates (e.g. helping an adult), see Supplementary Materials for details.

Why does the function space have higher fidelity for predicting human patterns of scene categorization? To concretize this result, we will examine a few failure cases for alternative features. Category names should reflect cognitively relevant categories, so what hurts the performance of the lexical distance model? This model considers the categories access road and road tunnel to have the lowest distance of all category pairs

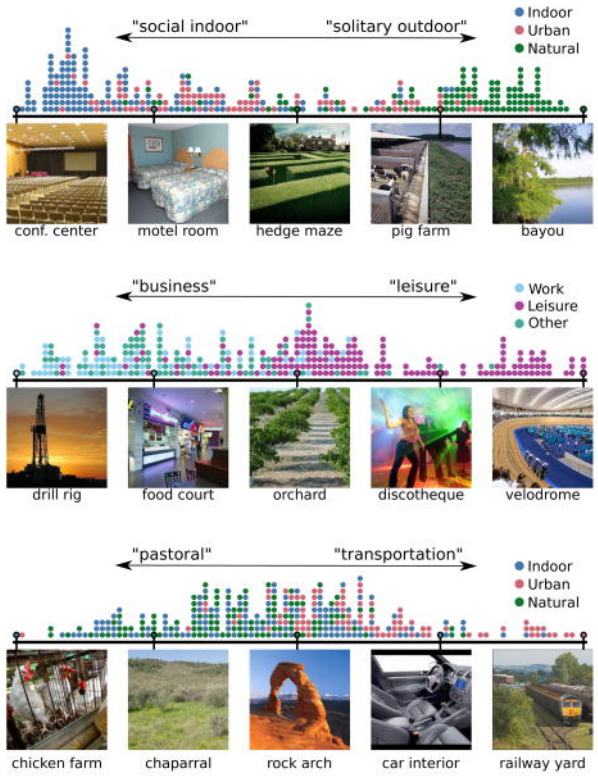


Figure 2.6: (Top): Distribution of superordinate-level scene categories along the first MDS dimension of the function distance matrix, which separates indoor scenes from natural scenes. Actions that were positively correlated with this component tend to be outdoor-related activities such as hiking while negatively correlated actions tend to reflect social activities such as eating and drinking. (Middle) The second dimension seems to distinguish environments for work from environments for leisure. Actions such as playing games are positively correlated while actions such as construction and extraction work are negatively correlated (Bottom). The third dimension distinguishes environments related to farming and food production (pastoral) from industrial scenes specifically related to transportation. Actions such as travel and vehicle repair are highly correlated with this dimension, while actions such as farming and food preparation are most negatively correlated.

(possibly because both contain the term road), while only 10% of human observers placed these into the same category. By contrast, the function model considered them to be rather distant, with only 35% overlap between functions (intersection over union). Shared functions included in transit / travelling and architecture and engineering work, while tunnels independently afforded rock climbing and caving and access roads often contained buildings, thus affording building grounds and maintenance work. If objects such as buildings can influence both functions and categories, then why dont objects fare better? Consider the categories underwater kelp forest and underwater swimming pool. The object model considers them to be very similar given the presence of water, but 80% of human observers consider them to be different. Similarly, these categories share only 17% overlap in functions, with the kelp forest affording actions such as science work, while the

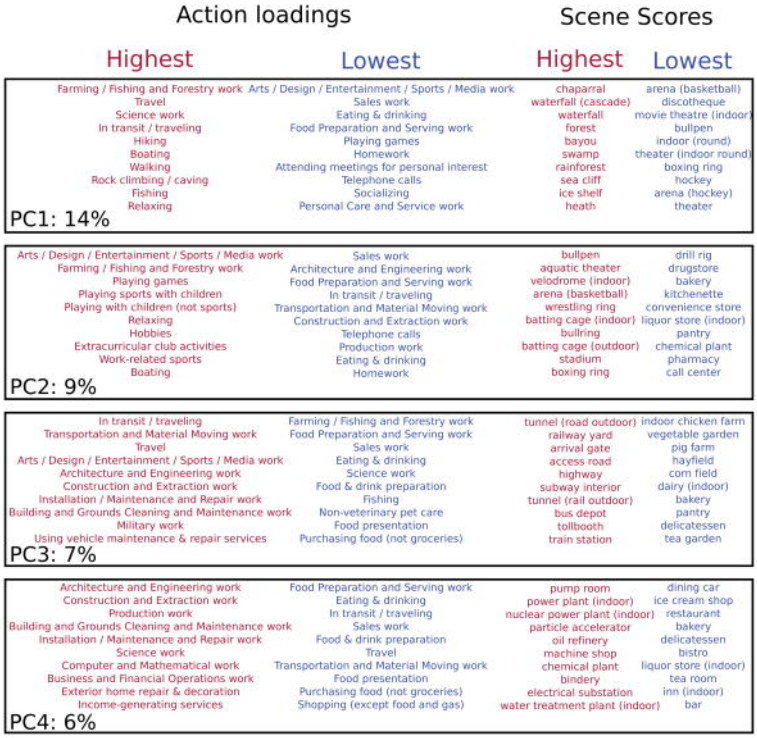


Figure 2.7: Principal components of function matrix. MDS was performed on the scene by function matrix, yielding a coordinate for each scene along each MDS dimension, as well as a correlation between each function and each dimension. The fraction of variance in scene distances explained by each dimension was also computed, showing that these first four dimensions capture 81% of the function distance model.

swimming pool affords playing sports with children.

Of course, certain failure cases of the function model should also be mentioned. For example, while all human observers agreed that bar and tea room were different categories, the function model considered them to be similar, given their shared functions of socializing, eating and drinking, food preparation and serving work etc. Similarly, the function model considered basketball arena and theatre to be similar, while human observers did not. Last, the function model also frequently confused scene categories that shared a particular sport, such as baseball field and indoor batting cage, while no human observers placed them in the same category. However, it should be noted that human observers also shared this last trait in other examples, with 55% of observers placing bullpen and pitchers mound into the same category.

2.1.3 Discussion

We have shown that human scene categorization is better explained by the action possibilities, or functions, of a scene than by the scenes visual features or objects. Furthermore, function-based features explained far more independent variance than did alternative models, as these models were correlated with human category

patterns only insofar as they were also correlated with the scenes functions. This suggests that a scenes functions contain essential information for categorization that is not captured by the scenes objects or visual features.

The current results cannot be explained by the smaller dimensionality of the function-based features, as further analysis revealed that function-based features outperformed other feature spaces using equivalent numbers of dimensions. Furthermore, this pattern was observed over a wide range of dimensions, suggesting that each functional feature contained more information about scene categories than each visual or object-based feature. Critically, the function-based model performed with similar fidelity on all types of scenes, which is a hallmark of human scene perception [205] that is not often captured in computational models. Indeed, indoor scene recognition is often much harder for computer models than other classification problems [333, 415] and this was true for our visual and object-based models, while the function model showed high fidelity for explaining indoor scene categorization.

The idea that the function of vision is for action has permeated the literature of visual perception, but it has been difficult to fully operationalize this idea for testing. Psychologists have long theorized that rapid and accurate environmental perception could be achieved by the explicit coding of an environments affordances, most notably in J.J. Gibsons influential theory of ecological perception [139]. This work is most often associated with the direct perception of affordances that reflect relatively simple motor patterns such as sitting or throwing. As the functions used in the current work often reflect higher-level, goal-directed actions, and because we are making no specific claims about the direct perception of these functions, we have opted not to use the term affordances here. Nonetheless, ideas from Gibsons ecological perception theory have inspired this work, and thus we consider our functions as conceptual extensions of Gibsons idea.

In our work, a scenes functions are those actions that one can imagine doing in the scene, rather than the activities that one reports as occurring in the scene. This distinguishes this work from that of activity recognition [3, 157, 454, 464], placing it closer to the ideas of Gibson and the school of ecological psychology.

Previous small-scale studies have found that environmental functions such as navigability are reflected in patterns of human categorization [151, 153], and are perceived very rapidly from images [150]. Our current results provide the first comprehensive, data-driven test of this hypothesis, using data from hundreds of scene categories and affordances. By leveraging the power of crowdsourcing, we were able to obtain both a large-scale similarity structure for visual scenes, but also normative ratings of functions for these scenes. Using hundreds of categories, thousands of observers and millions of observations, crowdsourcing allowed a scale of research previously unattainable. Previous research on scene function has also suffered from the lack of a comprehensive list of functions, relying instead on the free responses of human observers describing the actions that could be taken in scenes [151, 323]. By using an already comprehensive set of actions from the American Time Use Survey, we were able to see the full power of functions for predicting human categorization patterns. The current results speak only to categorization patterns obtained from unlimited viewing times, and future work will examine the extent to which function-based categorization holds for limited viewing times, similar to previous work [151, 150].

Given the relatively large proportion of variance independently explained by function-based features, we are left with the question of why this model outperforms the more classic models. By examining patterns of variance in the function by category matrix, we found that functions can be used to separate scenes along previously defined dimensions of scene variance, such as superordinate-level category [201, 260, 425], and between work and leisure activities [112]. Although the variance explained by function-based similarity does not come directly from visual features or the scenes objects, human observers must be able to apprehend these functions from the image somehow. It is therefore a question open for future work to understand the extent to which human observers bring non-visual knowledge to bear on this problem. Of course, it is possible that functions can be used in conjunction with other features for categorization, just as shape can be determined independently from shading [336], motion [202] or texture [138].

Some recent work has examined large-scale neural selectivity based on semantic similarity [187], or object-based similarity [397], finding that both types of conceptual structures can be found in the large-scale organization of human cortex. Our current work indeed shows sizeable correlations between these types of similarity structures and human behavioral similarity. However, we find that function-based similarity is a better predictor of behavior and may provide an even stronger grouping principle in the brain.

Despite the impressive predictive power of functions for explaining human scene categorization, many open questions are still left about the nature of functions. To what extent are they perceptual primitives as suggested by Gibson, and to what extent are they inherited from other diagnostic information? The substantial overlap between functions and objects and visual features (Figure 2.4B) implies that at least some functions are correlated with these features. Intuitively this makes sense as some functions, such as mailing may be strongly associated with objects such as a mailbox or an envelope. However, our results suggest that the mere presence of an associated object may not be enough: just because the kitchen supply store has pots and pans does not mean that one can cook there. The objects must conform in type, number, and spatial layout to jointly give rise to functions. Furthermore, some functions such as jury duty, waiting, and socializing are harder to associate with particular objects and features, and may require higher-level, non-visual knowledge. While the current results bypass the issue of how observers compute the functions, we must also examine how the functions can be understood directly from images in a bottom-up manner.

These results challenge many existing models of visual categorization that consider categories to be purely a function of shared visual features or objects. Just as the Aristotelian theory of concepts assumed that categories could be defined in terms of necessary and sufficient features, classical models of visual categorization have assumed that a scene category can be explained by necessary and sufficient objects [36, 397] or diagnostic visual features [345, 440]. However, just as the classical theory of concepts cannot account for important cognitive phenomena, the classical theory of scene categories cannot account for the fact that two scenes can share a category even when they do not share many features or objects. By contrast, the current results demonstrate that the possibility for action creates categories of environmental scenes. In other words, a kitchen is a kitchen because it is a space that affords cooking, not because it shares objects or other visual features with other kitchens.

Vocabulary Test for Potential AMT Participants

The categories included in the test were: bakery, volcano, highway, kitchen, restaurant, lighthouse, waiting room, forest, closet, sushi bar.

Finding Diagnostic Objects Associated with Functions

In order to understand the relationship between functions and objects, we performed a norming experiment on AMT. On each trial, participants were provided with a function name (hyperlinked to the ATUS description), and were asked to list the three objects they most associate with that action. Ten individual participants provided responses for each function, and 79 total workers contributed to the experiment. Participants qualified by (1) being based in the US; (2) having at least 2000 previously accepted hits on AMT with at least 98% acceptance and (3) having passed the scene vocabulary task described in the main text. The text responses were corrected for spelling and synonyms were standardized, and we computed the number of unique objects listed for each function. We reasoned that the fewer overall objects listed, the more agreement among individuals and thus the more the function is mapped to particular diagnostic objects. We split the 227 functions into two groups based on the number of objects listed. We then created two distance vectors based on these spaces and correlated these distances to the human distance vector. We found that the more object-based functions were correlated with the human distance vector at r=0.46 while the other group was correlated at r=0.42. To determine whether this correlation difference (0.032) is greater than what would be expected by taking random halves of functions, we simulated 10,000 random function splits and measured the correlation difference from these splits. The 95% confidence interval for this distribution was -0.09 to 0.09, indicating that our observed correlation difference of 0.032 was well within what could have been expected from chance alone.

List of Functions from American Time Use Survey

I. Personal care
Health related self-care
Sexual activity
Sleeping
Washing/dressing/grooming oneself

II. Household activities
Appliance repair & maintenance (self)
Building & repairing furniture
Cleaning home exterior
Email
Exercising & playing with animals

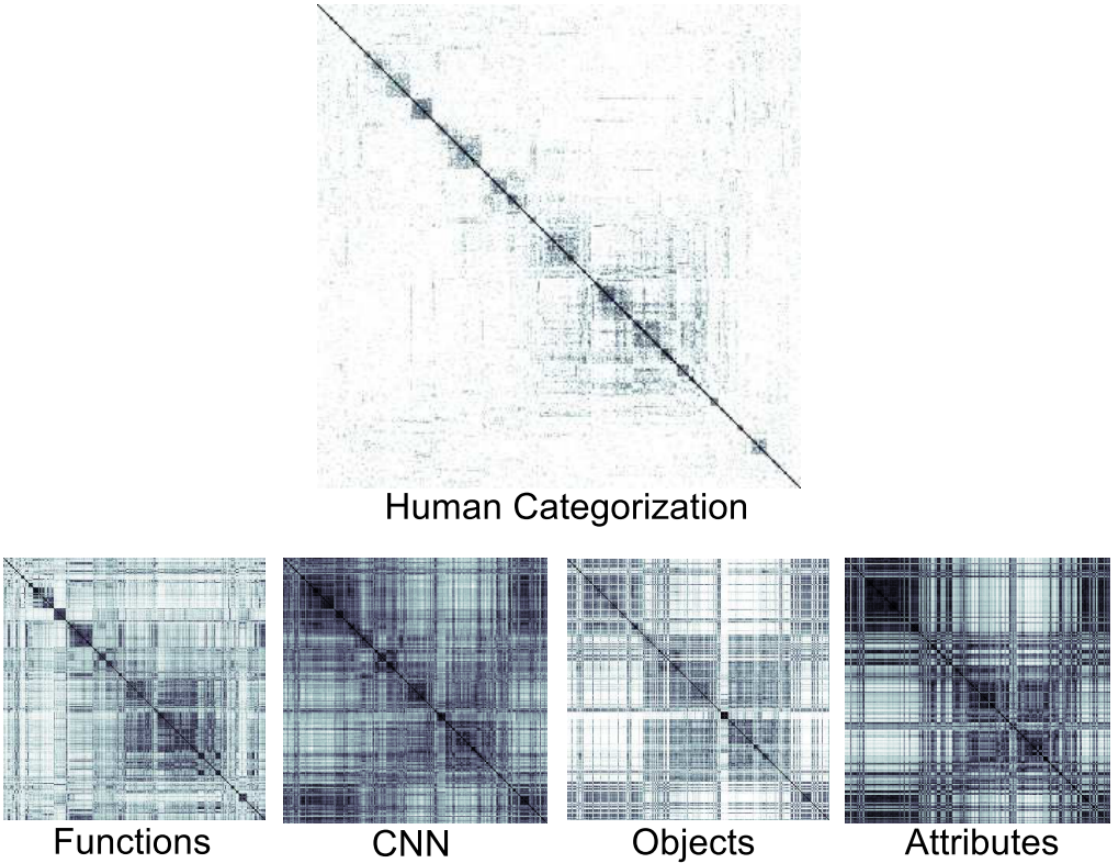


Figure 2.8: Distance matrices for top-four performing models, with human distance shown above for comparison (identical to Figure 2.3). All categories have been ordered according to the optimal leaf ordering for the human categorization data.

Exterior home repair & decoration

Financial management

Food & drink preparation

Food presentation

Grocery shopping

Home heating / cooling

Home security

Home-schooling children

Household organization & planning

Interior decoration & repair

Interior home cleaning

Kitchen & food clean-up
Laundry
Lawn/garden & plant care
Mailing
Maintaining home pool/pond/hot tub
Non-veterinary pet care
Sewing & repairing textiles
Storing household items
Vehicle repair & maintenance (self)

III. Caring for & helping household members

Arts & crafts with children
Attending childs events
Helping adult
Helping child with homework
Looking after adult
Looking after children
Obtaining medical care for adult
Obtaining medical care for child
Organizing & planning for adults
Organizing & planning for children
Physical care of adult
Physical care of children
Picking up / dropping off adult
Picking up / dropping off child
Playing sports with children
Playing with children (not sports)
Providing medical care to adult
Providing medical care to child
Reading with children
Talking with children

IV. Work & work-related activities

Architecture & engineering work
Arts / Design / Entertainment / Sports / Media work
Building and Grounds Cleaning and Maintenance work
Business and Financial Operations work

Community and social work
Computer and mathematical work
Construction and Extraction work
Education and library work
Farming / Fishing and Forestry work
Food Preparation and Serving work
Healthcare work
Income-generating hobbies & crafts
Income-generating performance
Income-generating rental property activity
Income-generating selling activities
Income-generating services
Installation / Maintenance and Repair work
Job interviewing
Job search activities
Legal work
Management/Executive work
Military work
Office and Administrative work
Personal Care and Service work
Production work
Protective services work
Sales work
Science work
Transportation and Material Moving work
Work-related eating/drinking
Work-related social activities
Work-related sports

V. Education

Attending school-related meetings & conferences
Education-related administrative activities
Extracurricular club activities
Homework
School music activities
Student government
Taking class for degree or certification

Taking class for personal interest

VI. Consumer purchases

Comparison shopping
Purchasing food (not groceries)
Purchasing gasoline
Shopping (except food and gas)

VII. Professional & personal care services

Banking
Buying & selling real estate
Out-of-home medical services
Using clothing repair & cleaning services
Using legal services
Using meal preparation services
Using other financial services
Using personal care services
Using professional photography services
Using vehicle maintenance & repair services
Using veterinary services

VIII. Household services

Using home repair & construction services
Using in-home medical services
Using interior home cleaning services
Using lawn & garden services
Using paid childcare services
Using pet services

IX. Government services & civic obligations

Civic obligations
Obtaining licenses & paying fees
Security screening
Using police & fire services
Using social services
Waiting

X. Eating & drinking

Eating & drinking

XI. Socializing, relaxing & leisure

Arts & crafts
Attending meetings for personal interest
Attending movies
Attending museums
Attending or hosting parties
Attending the performing arts
Collecting as a hobby
Computer use (not games)
Dancing
Gambling
Hobbies
Listening to music (not radio)
Listening to radio
Playing games
Reading for personal interest
Relaxing
Socializing
Tobacco use
Watching television & movies
Writing for personal interest

XII. Sports, exercise & recreation

Biking
Boating
Bowling
Camping
Doing aerobics
Doing gymnastics
Doing martial arts
Fencing
Fishing
Golfing
Hiking

Hunting
Participating in aquatic sports
Participating in equestrian sports
Participating in rodeo
Playing baseball
Playing basketball
Playing billiards
Playing football
Playing hockey
Playing racquet sports
Playing rugby
Playing soccer
Playing softball
Playing volleyball
Rock climbing / caving
Rollerblading / skateboarding
Running
Skiing / ice skating / snowboarding
Using cardiovascular equipment
Vehicle racing/touring
Walking
Watching aerobics
Watching aquatic sports
Watching biking
Watching billiards
Watching boating
Watching bowling
Watching dance
Watching equestrian sports
Watching fencing
Watching fishing
Watching golf
Watching gymnastics
Watching hockey
Watching live baseball
Watching live basketball
Watching live football

Watching live soccer
Watching live softball
Watching live vehicle racing
Watching martial arts
Watching people walk
Watching racquet sports
Watching rock climbing / caving
Watching rodeo
Watching rollerblading / skateboarding
Watching rugby
Watching running
Watching skiing / snowboarding
Watching volleyball
Watching weightlifting
Watching wrestling
Weightlifting
Working out
Wrestling
Yoga

XIII. Religious & spiritual activities

Attending religious services
Religious education
Religious practices

XIV. Volunteer activities

Volunteer at event
Volunteer work: attending meeting
Volunteer work: blood donation
Volunteer work: building
Volunteer work: clean up
Volunteer work: collecting goods
Volunteer work: computer use
Volunteer work: food preparation
Volunteer work: fundraising
Volunteer work: organizing
Volunteer work: performing

Volunteer work: providing care
Volunteer work: public safety
Volunteer work: reading
Volunteer work: teaching
Volunteer work: telephone calls
Volunteer work: writing

XV. Telephone calls

Telephone calls

XVI. Traveling

In transit / traveling
Travel
List of Functions from SUN Attribute Database
Sailing/boating
Driving
Biking
Transporting things or people
Sunbathing
Vacationing / touring
Hiking
Climbing
Camping
Reading
Studying / learning
Teaching / training
Research
Diving
Swimming
Bathing
Eating
Cleaning
Socializing
Congregating
Waiting in line / queuing
Competing
Sports

Exercise
Playing
Gaming
Spectating / being in an audience
Farming
Constructing / building
Shopping
Medical activity
Working
Using tools
Digging
Conducting business
Praying

2.2 Two Distinct Scene Processing Networks Connecting Vision and Memory

Natural scene perception has been shown to rely on a distributed set of cortical regions, including the parahippocampal place area (PPA) [116], retrosplenial cortex (RSC) [311], and the transverse occipital sulcus (TOS, aka the occipital place area, OPA) [166, 303]. More recent work has suggested that the picture is even more complicated, with PPA containing multiple subdivisions and the possible involvement of the parietal lobe [20]. Although there has been substantial progress in understanding the functional properties of each of these regions and the differences between them, the field has lacked a coherent overall framework for summarizing the overall architecture of the human scene processing system.

There is a long history of proposals for partitioning the visual system into separable components with different functions, such as spatial frequency channels [70], what versus where/how pathways [231, 288], or magnocellular, parcocellular, and koniocellular streams [208]. A division that is particularly relevant to natural scene perception is between the specific visual features present in the current glance of a scene, and the stable, high-level knowledge of where the place exists in the world, what has happened here in the past, and what possible actions we could take here in the future. For most cognitive and physical tasks we undertake in real-world places, the specific visual attributes we perceive are just a means to this end, of recalling and updating information about the physical environment; “the essential feature of a landmark is not its design, but the place it holds in a city’s memory” [301]. The connection between place and memory has been recognized for thousands of years, reflected in the ancient Greek method of loci that seeks to strengthen a memory by associating it with a physical location [466].

Some previous work has begun to point to this type of organizing principle among scene perception

regions. Mapping functional connectivity differences between pairs of scene-sensitive regions has revealed some consistent distinctions, with some regions more connected to visual cortex and others to parietal and medial temporal regions [20, 305]. Contrasting activity evoked by perceptual categorization tasks compared to semantic retrieval tasks shows a similar division between visual and higher-level cortex [124]. These experiments, however, have all been targeted, hypothesis-driven comparisons between regions with similar functional properties. It is unclear whether these divisions are major organizing principles of the brains connectivity networks, or simply subtle differences within a single coherent scene-processing network.

To answer this question, we took a data-driven approach to identifying scene-sensitive regions and clustering cortical connectivity. After applying a state-of-the-art connectivity algorithm [21] to generate spatially-coherent parcels based on high-resolution resting-state connectivity, we associate these parcels with components of the scene-processing network using category localizers, retinotopic field maps, category decoding, and a meta-analysis of previous work. We then perform hierarchical clustering and multidimensional scaling to show that there is a prominent, bilaterally symmetric division of scene-related regions into two separate networks: one includes TOS and the posterior portion of PPA (retinotopic maps PHC1 and PHC2), while the other is composed of the RSC, anterior PPA (aPPA), and the caudal inferior parietal lobule (cIPL). We show that the least well-known of these regions, the cIPL, actually has unique structural connectivity properties which makes it well suited to link visual perception with processing throughout the rest of the cortex.

Based on these results, as well as a review of previous studies, we propose that scene processing is fundamentally divided into two collaborating but distinct networks, with one focused on the visual features of a scene image and the other related to contextual retrieval and navigation. Under this framework, scene perception is less the function of a unified set of distributed neural machinery and more of "an ongoing dialogue between the material and symbolic aspects of the past and the continuously unfolding present" [19].

2.2.1 Materials and Methods

Imaging Data

The majority of the data used in this study was obtained from the Human Connectome Project (HCP), which provides detailed documentation on the experimental and acquisition parameters for these datasets [429]. We provide an overview of these datasets below.

Diffusion imaging data was used for the first 10 subjects from the January 2014 "Q3" HCP data release with complete data (subj ids 100408, 101915, 102816, 105216, 106016, 106319, 111009, 111514, 111716, 112819). Data were acquired using a multiband sequence at three different b-values (1000, 2000, 3000 s/mm2), with a total of 270 diffusion weighting directions and a resolution of 1.25mm isotropic.

The group-level functional connectivity data were derived from the 468-subject group-PCA eigenmaps, distributed with the June 2014 500 Subjects HCP data release. Resting-state fMRI data were acquired over four sessions (14 min, 33 seconds each) while subjects fixed on a bright cross-hair on a dark background,

using a multiband sequence to achieve a TR of 720ms at 2.0mm isotropic resolution (59412 surface vertices). These timecourses were cleaned using FMRIB's ICA-based Xnoiseifier (FIX) [368], and then the top 4500 eigenvectors for each voxel were estimated across all subjects using Group-PCA [392].

For the first 20 subjects within the "500 Subjects" release with complete data (and non-overlapping with the Q3 subjects: subj ids 101006, 101107, 101309, 102008, 102311, 103111, 104820, 105014, 106521, 107321, 107422, 108121, 108323, 108525, 108828, 109123, 109325, 111413, 113922, 120515), we created individual subject resting-state datasets by demeaning and concatenating their four resting-state sessions. We also obtained these subjects data from the HCP Working Memory experiment, in which they observed blocks of stimuli consisting of faces, places, tools, or body parts. We collapse across the two memory tasks being performed by participants (target-detection or 2-back detection).

To identify group-level scene localizers, we used data from a separate set of 24 subjects scanned at Stanford University (see below). Each subject viewed blocks of stimuli from six categories: child faces, adult faces, indoor scenes, outdoor scenes, objects (abstract sculptures with no semantic meaning), and scrambled objects. Functional data were acquired with an in-place resolution of 1.56mm, slice thickness of 3mm (with 1 mm gap), and a TR of 2s; a high-resolution (1mm isotropic) SPGR structural scan was also acquired to allow for transformation to MNI space. Full details of the localizer stimuli and acquisition parameters are given in our previous work [20].

Subjects

Scene localizer data was collected from 24 subjects (6 female, ages 22-32, including one of the authors). Subjects were in good health with no past history of psychiatric or neurological diseases, and with normal or corrected-to-normal vision. The experimental protocol was approved by the Institutional Review Board of Stanford University, and all subjects gave their written informed consent.

Resting-state Parcellation

We generated a voxel-level functional connectivity matrix by correlating the group-level eigenmaps for every pair of voxels and applying the arctangent function. We parcellated this 59412 by 59412 matrix into contiguous regions, using a generative probabilistic model [21]. This method finds a parcellation of the cortex such that the connectivity properties within each parcel are as uniform as possible, making multiple passes over the dataset to fine-tune the parcel borders. We set the scaling hyperparameter $\frac{2}{0} = 3000$ to produce a manageable number of parcels.

Scene localizers and retinotopic field maps

To identify PPA, RSC, and TOS, we deconvolved the localizer data from the 24 Stanford subjects using the standard block hemodynamic model in AFNI [96], with faces, scenes, objects, and scrambled objects as regressors. The Scenes $>$ Objects t-statistic was used to define PPA (top 300 voxels near the parahippocampal

gyrus), RSC (top 200 voxels near retrosplenial cortex), and TOS (top 200 voxels near the transverse occipital sulcus). The ROI masks were then transformed to MNI space, summed across all subjects, and mapped to the closest vertices on the group cortical surface. The cluster denoting highest overlap between subjects was then manually annotated.

A volumetric group-level probabilistic atlas [447] was used to define retinotopic field maps, by mapping each field map to the closest vertices on the group-level surface.

Scene category decoding

For each cortical parcel (generated from resting-state connectivity as described above), we measured its sensitivity to scenes versus other visual categories through a category decoding analysis. We first used a hemodynamic model to associate timepoints within the 20 HCP working memory datasets with specific stimulus categories. We labeled timepoints as corresponding to bodies, faces, places, or tools by constructing a boxcar timecourse denoting when each stimulus category was being displayed, convolving these indicators with the standard SPM hemodynamic response function provided with AFNI [96], rescaling the maximum value to 1, then re-thresholding to a binary indicator. Effectively, this produced a shift of the stimulus blocks by 5.55s to account for hemodynamic delay. The fMRI timecourses were cleaned by regressing out movement (6 degree-of-freedom translation/rotation and derivatives) and constant, linear, and quadratic trends from each run, then normalizing each voxel to have unit variance. Voxel timecourses were then averaged within each parcel, yielding a vector of average parcel activities for each timepoint.

Linear support vector machines (SVMs) were trained separately for each subject to discriminate scene timepoints from non-scene timepoints, and then tested on the other 19 subjects. We set the soft-margin hyperparameter c=1, but our results are not sensitive to this choice. Note that chance performance is 75%, since only 25% of the stimulus timepoints are scenes. Each subjects classifier assigned a weight to each parcel, indicating how strongly activity in this parcel predicted that a scene was being viewed. Parcels consistently assigned high positive weights were therefore most strongly associated with visual scene processing.

Meta-analysis

We sought to identify all fMRI studies involving scene memory, navigation, imagined experiences, or context memory that reported activation coordinates around the posterior parietal lobe. These coordinates were assumed to be in MNI space, unless identified as being in Talairach space, in which case we transformed the coordinates to MNI space [50]. Each coordinate was then mapped to the closest vertex on the group surface.

Parcel-to-parcel functional connectivity matrices

The 468-subject eigenmaps distributed by the HCP are approximately equal to performing a singular value decomposition on the concatenated timecourses of all 468 subjects, and then retaining the right singular values scaled by their eigenvalues [392]. This allows us to treat these eigenmaps as pseudo-timecourses, since dot

products (and thus correlations) between eigenmaps approximate the dot products between the original voxel timecourses. Given a parcellation, we computed the group-level connectivity between a pair of regions by taking the mean over all eigenmaps in each region, then correlating these mean eigenmaps and applying the Fisher z-transform (hyperbolic arctangent). We computed subject-level connectivity in the same way, using the resting-state timecourse for each voxel rather than the eigenmap.

Network Clustering and Multidimensional scaling

The 172 by 172 parcel functional connectivity matrix was converted into a distance matrix by subtracting every entry from the maximum entry. Ward clustering (unconstrained by parcel position) was used to compute a hard clustering into 10 networks. Separately, classical multidimensional scaling was also applied to the distance matrix, and the first three dimensions were used to assign voxels RGB colors (with each color channel scaled to span the full range of 0 to 255 along each axis) and to plot parcels in a 3D space. We performed the same operation on each subject-level matrix as well, and then aligned each subjects 3D pointcloud to the group pointcloud using a procrustes transform.

Structural connectivity

Probabilistic tractgraphy was performed on each of the 10 HCP diffusion datasets using FSL [194], by estimating up to 3 crossing fibers with bedpostx (using gradient nonlinearities and a rician noise model) and then running probtrackx2 using the default parameters and distance correction. 2000 fibers were generated for each of the $1.7x10^6$ white-matter voxels, yielding $3.4x10^9$ total sampled tracks per subject (approximately 34 billion tracks in total). We assigned each of the endpoints to gray-matter voxels using the 32k/hemisphere Conte69 registered standard mesh distributed for each subject, discarding the small number of tracks that did not have both endpoints in gray matter (e.g. cerebellar or spinal cord tracks). Since we are using distance correction, the weight of a track is set equal to its length.

The distance-based connectivity profile of a voxel was obtained by summing all of the voxels connections within 1cm bins based on Euclidean distance from the voxel. The profile for a parcel was then computed as the average of all its voxel profiles (rather than the sum, which does not control for differing parcel areas). Connectivity profiles for cIPL parcels vs. other parcels were compared using a two-way repeated measures ANOVA, with cIPL vs. other as the first factor and distance bin as the second factor.

We computed the structural connectivity between a pair of parcels A and B as the mean connectivity strength over all pairs of voxels with one voxel drawn from A and one drawn from B. Note that this also yields a measurement independent of parcel size.

2.2.2 Results

In order to reduce the complexity of the full 1.8-billion element whole-brain resting-state functional connectivity matrix, we first performed spatial parcellation using a generative modeling approach [21]. This

parcellation consisted of 172 spatially-coherent regions across both hemispheres, each of which contains voxels with near-uniform connectivity properties. The connectivity matrix between these 172 parcels captures more than 76% of the variance in the original connectivity matrix, despite being dramatically smaller (by five orders of magnitude). Representing the connectivity matrix in this way allows us to identify locations where functional connectivity profiles change rapidly (the boundaries between parcels), and lets us examine functional and connectivity properties at the more manageable and meaningful parcel level rather than at the voxel level.

Identifying Scene-Sensitive Parcels

Our first goal was to identify parcels that were related to processing visual scenes, using several different approaches as shown in Figure 7.1. Mapping group-level retinotopic field maps to the surface shows that the parcels exhibit an eccentricity-based organization (dividing foveal and peripheral voxels) in early visual areas, but that parcel boundaries begin to align with field map boundaries in later dorsal and ventral regions, as we have previously reported [21]. This alignment is especially prominent in parahippocampal regions PHC1 and PHC2, which are divided into anterior and posterior parcels. In the left (right) hemisphere, 86% (87%) of PHC1 voxels fall into the posterior parcel and 97% (72%) of PHC2 voxels fall into the anterior parcel. We also overlaid group-level localizer data (from a separate group of subjects) for scene-sensitive regions TOS, RSC, and PPA. TOS and RSC fall largely within single parcels (which we label the TOS and RSC parcels), while PPA runs perpendicular to parcel boundaries, extending through at least three separate parcels. The two posterior parcels correspond to PHC1 and PHC2 (which we collectively refer to as "posterior PPA", pPPA), and we label the most anterior parcel as "anterior PPA" (aPPA).

We can directly confirm that these parcels are scene-sensitive by applying our parcellation to task-fMRI data from the Human Connectome Project, and using the mean activity of each parcel as a feature for decoding scenes vs. other visual categories (faces, tools, bodies). These decoding accuracies were well above chance, even across subjects; a decoder trained on one subject could identify scene timepoints in other subjects with 85.1% accuracy (t_{19}=23.71, p<0.01; one-tailed t-test). Parcels that were consistently assigned positive weights for decoding scenes vs. other categories are identified in Figure 7.2. Scene-related parcels labeled from retinotopic maps and localizers exhibit high decoding weights (TOS: left t_{19}=3.95, p<0.01; right t_{19}=5.70, p<0.01; RSC: left t_{19}=4.95, p<0.01; right t_{19}=2.80, p<0.01; PHC1: left t_{19}=3.83, p<0.01; right t_{19}=1.06, n.s.; PHC2: left t_{19}=4.95, p<0.01; right t_{19}=5.66, p<0.01; aPPA: left t_{19}=1.73, p<0.05; right t_{19}=7.34, p<0.01; one-tailed t-test).

Interestingly, scene selectivity extends dorsally beyond TOS, into the caudal inferior parietal lobule (cIPL). Labeling the three parcels in this region cIPL1-3 (ordered posterior to anterior along the angular gyrus), both cIPL1 and cIPL2 consistently show discriminative weights for the (unfamiliar) localizer scenes (cIPL1: left t_{19}=9.61, p<0.01; right t_{19}=8.34, p<0.01; cIPL2: left t_{19}=3.87, p<0.01; right t_{19}=3.58, p<0.01) while cIPL3 does not (left t_{19}=-1.16, n.s; right t_{19}=1.48, n.s.). This result suggests that there may be scene-related activity anterior to typically-defined TOS, but does not provide clear evidence for a separate region

with different functional properties. Scene localizers, however, are missing a critical component of real-world scene perception; since they typically include only unfamiliar scenes, they may fail to robustly activate memory and contextual networks engaged in processing familiar environments. A meta-analysis of previous studies shows that personally familiar places robustly activate cIPL, especially around cIPL2 and cIPL3 (Figure 7.3). This activation appears for a wide variety of tasks, including memory for visual scene images [115, 117, 293, 416, 426], learning navigational routes [49, 57], and even simply imagining past events or future events in familiar places [164, 413]. This same region can also be activated by recalling non-place stimuli (including words and objects), if the stimuli are associated with strong memory of the source context [198, 329, 437]. These studies, along with our previous work showing connectivity differences between TOS and cIPL [20], provide strong evidence that the caudal inferior parietal lobe is in fact a separate, important component of the scene-processing system.

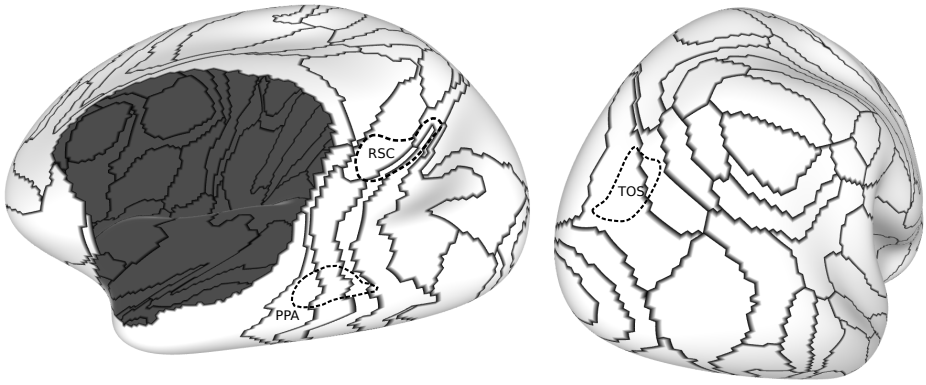


Figure 2.9: **Relationship between resting-state parcels, retinotipic maps, and scene localizers**. Group-level visual field maps and functional localizers are overlaid on parcels derived from resting-state connectivity patterns (black borders). RSC and TOS largely fall within a single parcel, with TOS corresponding roughly to V3B. Ventrally, PHC1 and PHC2 are well divided into two separate parcels, with PPA extending anteriorly into a parcel we denote aPPA.

Clustering Parcels into Networks

Having identified these eight (bilateral) parcels critical to scene perception, we clustered the whole-brain connectivity matrix to identify 10 functionally-connected networks. This data-driven analysis groups together parcels that all have high functional connectivity with one another, regardless of their spatial position.As shown in Figure 7.4, these networks are remarkably symmetric between hemispheres, and split scene perception regions into two separate categories. Posterior parcels - TOS, cIPL1, PHC1, and PHC2 - were clustered into visual network (dark blue) covering all of visual cortex outside of the early foveal cluster. Anterior parcels - cIPL2, cIPL3, RSC, and aPPA - were clustered into a separate parietal/medial-temporal network (pink), which also included anterior temporal and medial frontal parcels. This corresponds to a portion of the known default mode regions, with other default mode regions being grouped into a separate network

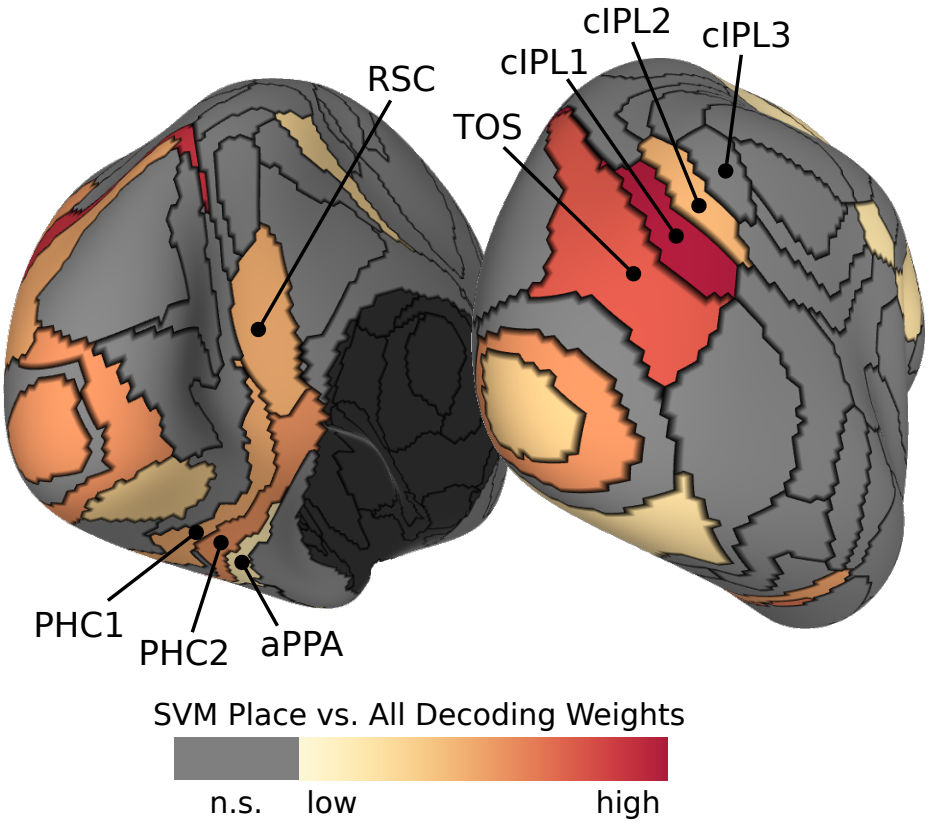


Figure 2.10: **Parcel scene decoding weights**. Linear SVMs were trained to classify unfamiliar scenes vs other images (faces, tools, bodies) based on mean activity in each resting-state parcel. Colored regions are those having significant positive weights across subjects ($p<0.05$). High activity in the parcels identified using field maps and scene localizers (Figure 1) predict that subjects are viewing scenes, and these positive weights extend from TOS partially onto the angular gyrus.

(green). The dividing line between the visual and context networks falls consistently near the edge of known retinotopic maps, suggesting a division between regions strongly tied to the current retinal input and those which are more driven by internally-driven processes and integrate information over longer time-scales. If the number of clusters is increased, divisions within these networks appear, first between TOS and pPPA, and then between RSC/cIPL and aPPA.

Rather than performing a hard clustering into distinct groups, we can use classical multidimensional scaling (MDS) to embed parcels into a three-dimensional space. Distances in this space approximate the functional connectivity strength between parcels, such that strongly-connected parcels are close together. Setting the RGB color of each parcel based on its position in this three-dimensional embedding space gives a soft clustering (Figure 7.5(a)). Moving along either the dorsal (TOS-cIPL) or ventral (PHC-aPPA) boundaries between scene regions produces rapid changes in functional connectivity properties, visualized in embedding space in Figure 7.5(b-c). In both cases, the most posterior regions (TOS and PHC1) show strong connectivity

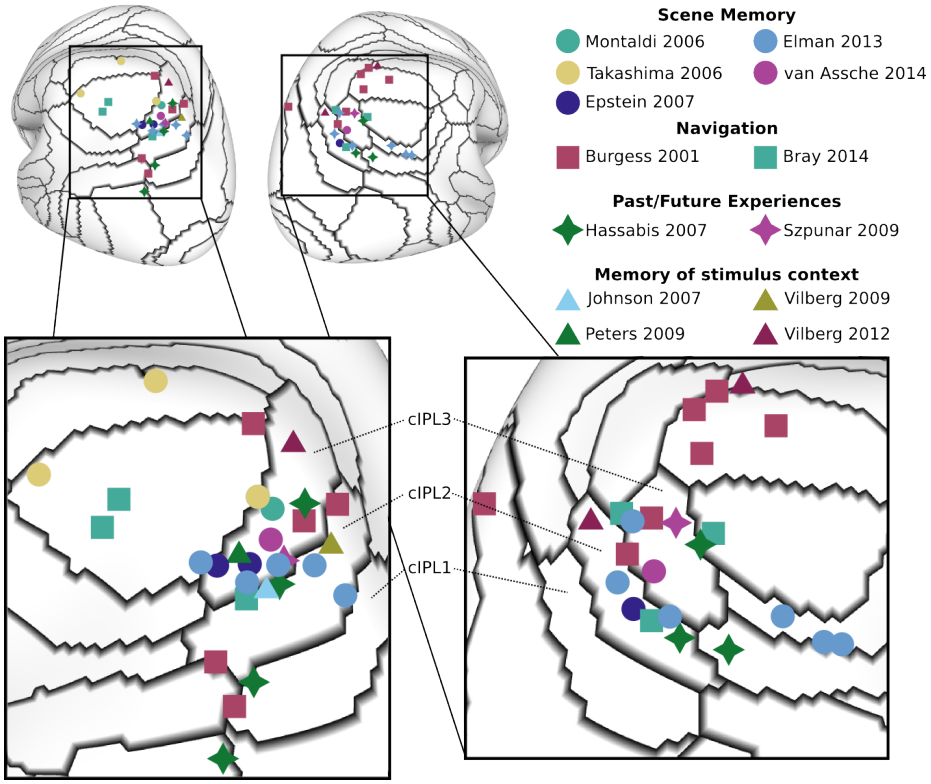


Figure 2.11: **Meta-analysis of cIPL involvement in place memory**. Although not typically identified as a scene-sensitive region, the posterior parietal lobe is consistently activated in studies involving familiar places. Perceiving images of familiar scenes, learning navigational routes, or imagining events in familiar places produces activation clustered around cIPL2-3. This same region also appears in memory studies of non-scene stimuli associated with a strong context.

to other parcels in visual cortex, while the most anterior regions (cIPL3 and aPPA) are instead more related to default mode regions. To statistically evaluate this difference, we measure the connectivity between each scene-related parcel and a default-mode reference parcel on the opposite side of cortex (medial versus lateral), to avoid spurious connectivity due to local noise correlations. For the dorsal parcels, we measure connectivity to RSC, and for the ventral parcels, we measure connectivity to cIPL3. Along the dorsal boundary, we see significant increases in connectivity to RSC when moving from TOS to cIPL1 (Left: t_{19}=6.98, p<0.01; Right: t_{19}=6.35, p<0.01; two-tailed paired t-test), from cIPL1 to cIPL2 (Left: t_{19}=7.72, p<0.01; Right: t_{19}=6.16, p<0.01), and from cIPL2 to cIPL3 (Right: t_{19}=2.44, p<0.05). We observe a similar (though less dramatic) increase in connectivity to cIPL3 when moving from PHC1 to PHC2 (Left: t_{19}=4.21, p<0.01; Right: t_{19}=2.68, p<0.05) and PHC2 to aPPA (Right: t_{19}=3.03, p<0.01). These results (Figure 7.5(d-e)) indicate that the borders between the visual and context networks are not artifacts of the clustering procedure, but are in fact marked by rapid changes in connectivity properties.

Given the dramatic differences in functional connectivity properties among the scene parcels (especially

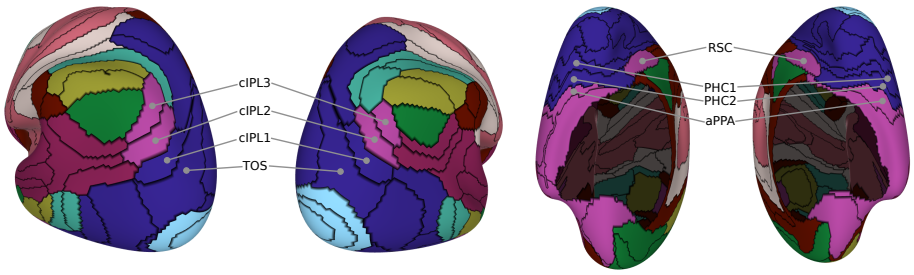


Figure 2.12: **Connectivity clustering of parcels**. Performing hierarchical clustering on the resting-state parcels based on their pairwise functional connectivity reveals that the scene processing network is split across two networks: a visual network (blue) which includes TOS and PHC1/2, and a parietal/medial-temporal network including cIPL, RSC, and aPPA. The visual network covers known retintopic field maps outside the early fovea, while the parietal/medial-temporal network corresponds to a portion of the default mode network.

cIPL, e.g. in Figure 7.5(d)), we examined whether these regions also differed in terms of structural connectivity, using diffusion imaging. We sampled 34 billion white matter seed locations across 10 subjects, and performed probabilistic tractography to identify the likely endpoints of the fiber tract passing through that seed. As shown in Figure 7.6, the cIPL parcels were qualitatively different from all other scene parcels, with both higher overall fiber incidence (per unit area) and a disproportionate number of long-range fibers (cIPL parcels vs. others, $F_{1,9}=191.24$, $p<0.01$; distance bin, $F_{19,171}=47.04$, $p<0.01$; interaction, $F_{19,171}=14.82$, $p<0.01$). These connections are widely distributed over posterior parietal, lateral and medial temporal, and prefrontal cortices, indicating the cIPL is structurally well-positioned to connect visual scene information with a wide variety of other cortical networks.

2.2.3 Discussion

By combining a variety of data sources including function and structural connectivity data, task-fMRI, retinotopic maps, and a meta-analysis of previous results we have shown converging evidence for a functional division of scene-processing regions into two separate networks (summarized in Figure 7.7). The visual network covers retintopically-organized regions including TOS and posterior PPA (pPPA), while a separate memory-related network connects cIPL, RSC, and anterior PPA (aPPA). This division emerges from a purely data-driven network clustering, suggesting that this is a core organizing principle of the visual system. Our data also support a much more prominent role for cIPL in processing real-world familiar scenes, since it is well positioned both functionally and structurally to connect scene processing with the rest of the brain.

Subdivisions of the PPA

The division of the PPA into multiple anterior-posterior subregions with differing connectivity properties replicates our previous work (Baldassano et al., 2013) on an entirely different large-scale dataset, and shows that there is a strong connection between connectivity changes in PPA and the boundaries of retinotopic field

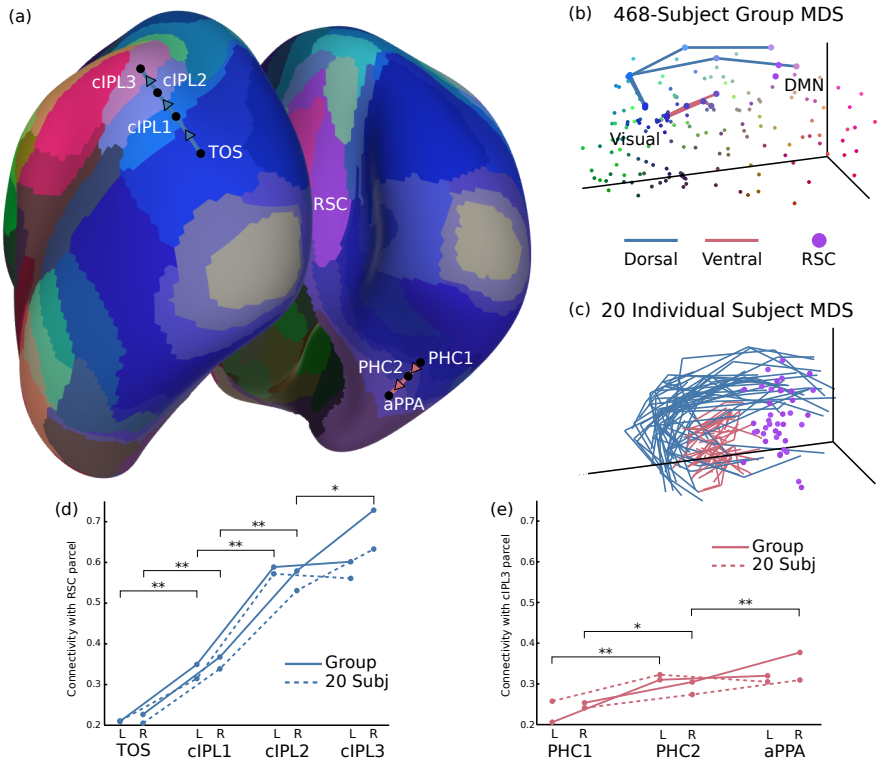


Figure 2.13: **Connectivity changes across the network border**. (a) Rather than performing a hard clustering assignment as in Figure 7.4, we can perform classical MDS on the parcel connectivity network and set regions RGB values based on their positions in a three-dimensional embedding space. This shows a similar result to hierarchical clustering, with abrupt connectivity changes across scene networks. (b) In MDS space, moving dorsally from TOS to cIPL3 produces the curves shown in blue, while moving ventrally from PHC1 to aPPA produces the curves shown in red. These curves move in parallel out of the retinotopic cluster toward the default mode cluster. (c) Plotting these curves for 20 individual subjects shows a similar pattern in each subject, with curves moving in parallel toward RSC (purple dots). (d) The connectivity between scene parcels and RSC increases dramatically as we move dorsally from TOS to cIPL3. (e) Connectivity with cIPL changes more subtly but significantly when moving ventrally from PHC1 to aPPA. *,** $p<0.05$, $p<0.01$

maps. There is now a growing literature on anterior versus posterior PPA, including not only connectivity differences [305] but also the response to low-level [306] and high-level [253, 320] scene properties. Our results place this division into a larger context, and demonstrate that the connectivity differences within PPA are not just an isolated property of this region but a general organizing principle for scene-processing regions.

This subdivision may be the key to resolving a long-standing debate over the role of context effects in PPA. Some have proposed that PPA is primarily driven not by scenes per se but any stimuli with strong spatial contextual associations [10], and that these associations drive activity during even the early stages of perception [238]. Others have argued that PPA is only involved in visual spatial layout processing, and that

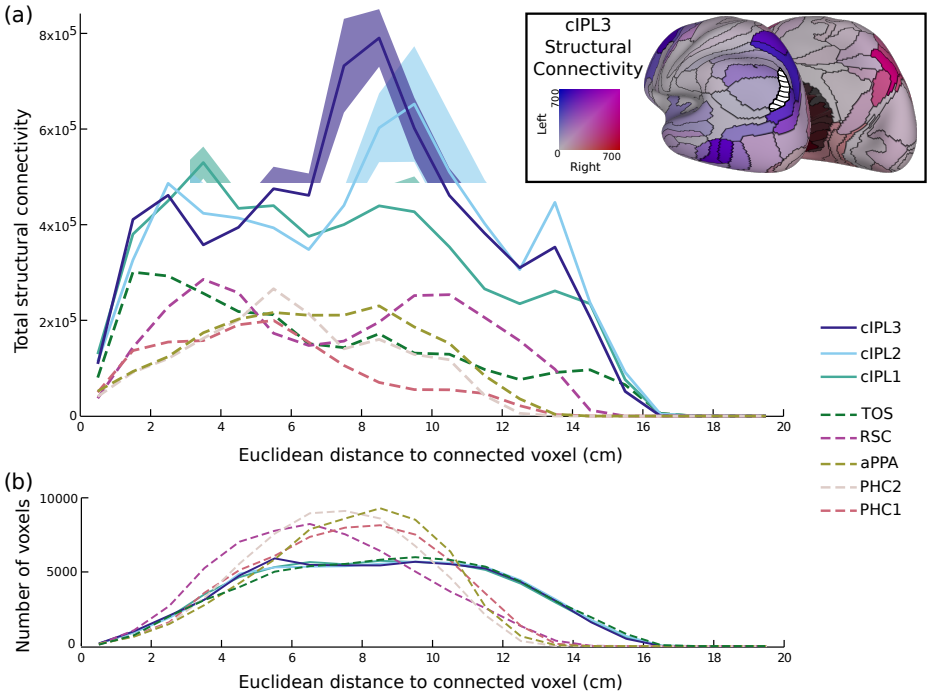


Figure 2.14: **Structural connectivity profiles of scene parcels**. (a) The connectivity between voxels in each parcel and the rest of the brain is plotted as a function of Euclidean distance (averaged between hemispheres, shaded regions show standard error of the mean). The cIPL parcels shows a distinct profile, both in overall connectivity strength and an emphasis on long-range connectivity. As shown in the inset, cIPL3 is structurally connected to a distributed set of cortical regions (primarily restricted to the same hemisphere). (b) The peak of cIPL connectivity around 10 cm is not driven by simple geometry, since the percentage of the cortex that is this distance away from cIPL is smaller than for other parcels such as RSC and those in PPA.

context effects are mostly an artifact of later imagery [120]. We argue that both these descriptions may be correct, but for different portions of PPA, with pPPA more related to concrete features of a visual scene and aPPA more related to general spatial context. In fact, the maps illustrated in these papers (Figure 4 in [10]; Figure 4 in [120]) suggest this type of anterior/posterior division.

The visual network

The visual network shows a close correspondence with the full set of retinotopic maps identified in previous studies [51, 186, 447], extending through the intraparietal sulcus (IPS) and laterally to hMT+. Our observation that TOS overlaps at the group level with retinotopic maps, primarily V3B, is consistent with prior measurements made in individual subjects [33, 304]. The only portion of cortex with known retinotopic maps that is not clustered in this network is the shared foveal representation of early visual areas, which segregates into its own cluster. One possible explanation is that our connectivity measures are based on eyes-open resting-state scans, during which a subjects fovea is being stimulated with a bright cross. This

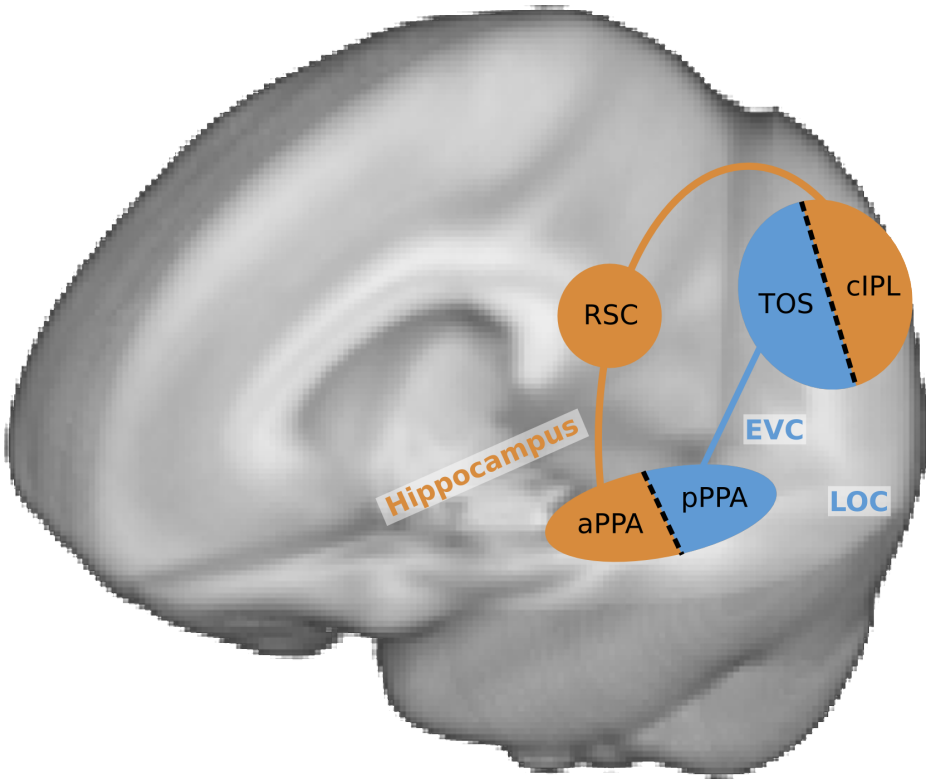


Figure 2.15: **Two-network model of scene perception**. Our results provide strong evidence for dividing scene-sensitive regions into two separate networks. TOS and posterior PPA (PHC1/2) process the current visual features of a scene (in concert with other visual areas, such early visual cortex and LOC), while cIPL, RSC, and anterior PPA perform higher-level context and navigation tasks (drawing on long-term memory structures such as the hippocampus).

stimulation may be the dominant signal in this region, resulting in a suppression of the intrinsic fluctuations used to define resting-state networks.

TOS and posterior PPA have been shown to be responsive primarily to visual features of a stimulus, rather than higher-level attributes such as familiarity. Posterior PPA has a preferential response to high spatial frequencies [335], and both posterior PPA and TOS are activated by rectilinear shapes [306], even in non-scene images. Also, neither TOS nor posterior PPA show reliable familiarity effects ([117], but see further discussion below).

The functional distinction between pPPA and TOS is currently unclear. Previous work has speculated about the purpose of the apparent ventral and dorsal duplication of regions sensitive to large landmarks, proposing that it may be related to different output goals (e.g. action planning in TOS, object recognition in pPPA) [226], or to different input connections (e.g. lower visual field processing in TOS, upper visual field processing in pPPA) [232].

The context and navigation network

The network of parahippocampal, retrosplenial, and posterior parietal regions we identify has been emerged independently in many different fields of neuroimaging, outside of scene perception. Meta-analyses of internally-directed tasks such as theory of mind, autobiographical memory, and prospection have identified this as a core, re-occurring network [214, 395] (and component C10 of [467])). This network also appears in navigation [57, 393], recalling the study context of a stimulus [168, 198, 329, 427], recognition of personally familiar locations[115, 426], viewing objects with strong contextual associations [11], and thinking about past or imagined events in familiar contexts [164, 413, 414].

The broad set of tasks which recruit this network have been summarized in various ways, such as "scene construction" [165], "mnemonic scene construction" [13], or "relational processing" [113]. A review of memory studies referred to this network as the posterior medial (PM) memory system, and proposed that it is involved in any task requiring "situation models" relating entities, actions, and outcomes [340].

Sometimes this network appears as part of the larger default mode network, which includes other regions such as parts of medial prefrontal cortex. However, the functional and anatomical structure of the default mode network suggests that it not a single coherent structure, and that the parietal/medial-temporal portion is in fact a distinct subnetwork [13, 14, 468].

The specific functions of the individual components of this network have also been studied in a number of contexts. RSC appears to be most directly involved in orienting the viewer to the structure of the environment (both within and beyond the borders of the presented image) for the purpose of navigational planning; it encodes both absolute location and facing direction [119, 270, 433], integrates across views presented in a panoramic sequence [321], and shows strong familiarity effects [117, 118]. This is consistent with rodent neurophysiological studies, which have identified head direction cells in this region [83]. RSC is not sensitive to low-level rectilinear features in non-scene images such as objects or textures, though it does show some preference for rectilinear features in images of 3D scenes [306].

Anterior PPA has been less well-studied, since it was not recognized as a separate region within the PPA until recently, but has been most strongly associated with coding the size of a scene [320]. Its representation of scene spaciousness draws on prior knowledge about the typical size of different scene categories, since it is affected by the presence of diagnostic objects [253].

The cIPL (also referred to as pIPL, PGp, or the angular gyrus) has been proposed as a "cross-modal hub" [14] that connects visual information with other sensory modalities as well as knowledge of the past. It is more intimately associated with visual cortex than most lateral parietal regions, since it has strong anatomical connections to higher-level visual regions in humans and macaques [74], and has a neurotransmitter receptor distribution similar to V3v and distinct from the rest of the IPL [75]. It is primarily involved in two related kinds of tasks. First, it supports contextual recall, showing both increases in mean activity [293, 437] as well as voxel-level activity patterns related to the specific context associated with an item [234]. Second, it performs temporal integration, sustaining activity under long delay periods [438], and accumulating both visual and auditory information over long time-scales [249]. Consistent with our structural connectivity results, its

functional connections are distributed and flexible, coupling to the dorsal attention network during a spatial learning task [49] or to dorsolateral prefrontal and extrastriate visual cortex during successful recollection [217]. Based on these properties, it has been proposed [436] that this region implements the multi-modal episodic buffer proposed by [18].

Given cIPLs involvement in a diverse set of tasks, it has not traditionally been identified as a central part of the scene perception system. However, our results suggest a deep connection between cIPL and understanding real-world places, which (unlike typical localizer images) are associated with a wealth of memory, context, and navigational information. Our meta-analysis shows that cIPL is selectively responsive to familiar scenes (arguably the most common high-context stimuli in everyday life), but this property has largely gone unnoticed in the scene perception literature; for example, one of the studies in Figure 7.3 showing cIPL activation [117] described this location only as "near TOS." More importantly, our clustering analyses revealed that cIPL is tightly coupled (at rest) with RSC and aPPA, two regions that are widely recognized as performing scene-specific processing. Lesion studies support this view that the posterior parietal lobe is primarily involved in scene-related functions (such as orienting to a previously learned map based on the current view), since these abilities can be selectively impacted without general memory deficits (reviewed in [231]).

Contrasting the two networks

Although our work is the first to propose the visual versus context networks as a general framework for scene perception, several previous studies have shown differential effects within these two networks. Contrasting the functional connectivity patterns of RSC vs. TOS or LOC [305] or anterior vs. posterior PPA [20] show a division between the two networks, consistent with our results. Contrasting scene-specific activity with general (image or word) memory retrieval showed an anterior vs. posterior distinction in PPA and cIPL/TOS, with only more anterior regions (aPPA and cIPL, along with RSC) responding to content-independent retrieval tasks [124, 198]. Our two-network division is also consistent with the dual intertwined rings model, which argues for a high-level division of cortex into a sensory ring and an association ring, the second of which is distributed but connected into a continuous ring through fiber tracts [280].

Open questions

The anterior/posterior pairing of aPPA/pPPA and cIPL/TOS raises the question of whether there is a similar anterior/posterior division in RSC. There is some evidence to suggest that this is the case: wide-field retinotopic mapping using natural scenes shows a partial retinotopic organization in RSC [186], and RSCs response to visual rectilinear features appears to be limited to the posterior portion [306]. However, we did not observe strong scene-selective responses in neighboring parcels near RSC (see Figure 7.2), a study of retinotopic coding in scene-selective regions failed to find any consistent topographic organization to RSC responses [452], and previous analyses of the functional properties of anterior versus posterior RSC have not found any significant differences [320].

Another interesting question is how spatial reference frames differ between and within the two networks. Given its retinotopic fieldmaps, the visual network presumably represents scene information relative to the current eye position; previous work has argued that this reference frame is truly retina-centered and not egocentric [148, 452]. The context network, however, likely transforms information between multiple reference frames. Models of spatial memory suggest that medial temporal lobe (possibly including aPPA) utilizes an allocentric representation, while the posterior parietal lobe (possibly including cIPL) is based on an egocentric reference frame, and that the two are connected via a transformation circuit in RSC that combines allocentric location and head direction [66, 432]. There is some recent evidence for this model in human neuroimaging: posterior parietal cortex codes the direction of attention in an egocentric reference frame (even for positions outside the field of view) [375], and RSC contains both position and head direction information (anchored to the local environment) [270]. This raises the possibility that another critical role of cIPL could be to transform retinotopic visual information into a stable egocentric scene over the course of multiple eye movements. The properties of aPPA, however, are much less clear; it seems unlikely that it would utilize an entirely different coordinate system than neighboring PHC1/2, and some aspects of the scene encoded in aPPA (such as overall scene size [320]) dont seem tied to any particular coordinate system.

Conclusion

Based on a review of previous literature, as well as novel comparisons of scene-related regions with data-driven clustering analyses, we have proposed a unifying framework for understanding the neural systems involved in processing both visual and non-visual properties of natural scenes. This new two-network classification system makes explicit the relationships between known scene-sensitive regions, re-emphasizes the importance of the functional subdivision within the PPA, and incorporates posterior parietal cortex as a primary component of the scene-understanding system. Our proposal, that much of the scene-processing network relates more to contextual and navigational information than to specific visual features, suggests that experiments with unfamiliar natural scene images will give only a partial picture of the neural processes evoked in real-world places. Experiencing our visual environment requires a dynamic cooperation between distinct cortical systems, to extract information from the current view of a scene and then integrate it with our understanding of the world and determine our place in it.

2.2.4 Acknowledgements

Funding was provided by a National Science Foundation Graduate Research Fellowship under grant number DGE-0645962, and Office of Naval Research Multidisciplinary University Research Initiative grant number N000141410671. Data were provided in part by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. We thank the Richard M. Lucas Center for Imaging, the Center for Cognitive and Neurobiological Imaging, and Michael Arcaro for helpful discussions.

2.3 On the Technology Prospects and Investment Opportunities for Scalable Neuroscience

Two major initiatives to accelerate research in the brain sciences have focused attention on developing a new generation of scientific instruments for neuroscience. These instruments will be used to record static (structural) and dynamic (behavioral) information at unprecedented spatial and temporal resolution and report out that information in a form suitable for computational analysis. We distinguish between recording — taking measurements of individual cells and the extracellular matrix — and reporting — transcoding, packaging and transmitting the resulting information for subsequent analysis — as these represent very different challenges as we scale the relevant technologies to support simultaneously tracking the many neurons that comprise neural circuits of interest. We investigate a diverse set of technologies with the purpose of anticipating their development over the span of the next 10 years and categorizing their impact in terms of short-term [1-2 years], medium-term [2-5 years] and longer-term [5-10 years] deliverables.

- short-term options [1–2 years] — The most powerful recording and reporting technologies currently available all use some form of imaging in which some portion of the acoustic or electromagnetic spectrum is used to illuminate a target tissue and the resulting response, attenuated by absorption, reflection and scattering, is analyzed to extract useful information about structure, e.g., cytoarchitectural details, and function, e.g., membrane potentials and spike-timing data. These relatively mature technologies largely finesse the problems relating to powering reporting devices and carrying out the required computations involved in signal processing, compression and transmission by performing all these functions external to the brain. Example technologies include electroencephalography, focused ultrasound, magnetic resonance imaging, microendoscopy, photoacoustic imaging, two-photon calcium imaging, array tomography for proteomics, immunoflorescence for genomics and light-sheet fluorescence microscopy. This class of technologies also includes our current best non-invasive options for the study of human subjects.

  Incremental improvements in these technologies are likely to continue unabated for some time, enabled by advances in biology and physics and funded and motivated by applications in medicine and materials science. In order to better resolve features of interest, biochemists are developing new reagents that are differentially absorbed by cellular structures and serve to alter the local spectral characteristics of illuminated tissue. Tissue samples can be prepared in such a way that structures that would normally absorb or scatter light such as the bilipid layers that comprise cell membranes are rendered transparent. Dyes can be integrated into living tissue and used as indicators for the presence of molecules of interest or to measure the observable state of cellular processes such as changes in membrane potential. Of course the addition of foreign molecules alters the optical properties of the tissue limiting penetration depth. Resolution is limited by light scattering and the resulting loss in penetration depth this causes. Advances in molecular functional magnetic resonance imaging (fMRI) may ultimately allow us to combine the specificity of electrophysiological recording techniques with the noninvasiveness

and whole-brain coverage of current fMRI technology.

- medium-term options [2–5 years] — Biological organisms demonstrate a wide variety of molecules that carry out cellular computing, sensing and signalling. These biomolecular devices are several orders of magnitude more efficient than CMOS devices. Efforts so far to harness biological machines to perform logic and arithmetic are hampered by the fact that biological circuits coerced into implementing these traditional computing operations are orders of magnitude slower their CMOS counterparts. However, for those computations required for survival, natural selection has stumbled on some incredibly efficient solutions. Bioengineers are compiling libraries of biomolecules found in nature that perform such environmentally-specialized computations. It is often said that, if you need a specific functional component for manipulating molecular information, you just have to find an organism that requires that function and re-purpose it for your application. This synthetic-biology approach may be our most expedient option for getting to the next level in the next 2-5 years. Examples currently under development include using retroviruses such as "tame" variants of the rabies and HIV virus to trace circuits, using DNA polymerase, the enzyme responsible for DNA replication, to record electrical potentials, using DNA sequences to barcode cells and cell products and modern, high-speed genome sequencing technology to create a map of all the synaptic connections in sample of tissue. While it seems likely there will be proof-of-concept demonstrations of such technologies in the next few years, it will be some time before they develop to a point where they can be applied routinely in animal studies, and longer before they can safely be used in human studies.

- longer-term options [5–10 years] — Some neuroscientists believe that the ability to observe neural activity at substantially higher spatial and temporal resolution than currently possible will lead to new discoveries and new approaches to neuroscience. One approach to achieving this level of detail is to enlist bioengineers and nanotechnologists to develop nanoscale recording and reporting devices that can be collocated with targets of interest in the neural tissue. There are number of challenges to achieving such technology. Moore's law and related predictions of progress in miniaturizing integrated circuits suggest that in the next five years or so we will be able to manufacture chips of roughly the same size as a single cell — less than 10 μm — providing on the order of 10,000 transistors. We would also have to reduce their power requirements to less than 10 nw to have some chance of powering the devices and dissipating their waste heat without causing cellular damage. Most semiconductors used in chips are toxic to cells and so we would have to develop alternative technologies like silicon bicarbide or find better ways of chemically isolating existing technologies such as cadmium. Perhaps the biggest challenge involves solving the related reporting problem: getting the information out of the brain; obvious approaches to utilizing existing RF or optical communication technologies do not scale to billions of nanoscale reporters. Safe, scalable solutions in this arena will likely require fundamental advances in science and engineering to achieve.

Each of these three planning horizons, 1–2, 2–5, 5–10 years, offers opportunity for investment. In 1–2 years,

incremental improvements in functional magnetic resonance will continue accelerate the study of cognitive disorders of development and aging. Less-invasive ultrasound-based technologies for stimulation and surgical intervention are poised to deliver new treatments for neurodegenerative disease. In the 2–5 year time frame, advances in synthetic-biology, DNA sequencing and multi-cellular recording promise insights into the function of neural circuits involving thousands of cells. Such insights are likely to yield new approaches to efficient computing, improved implants and prosthetic devices, and methods of harnessing natural computation. In 5-10 years, nanoscale sensors and robotics devices and the development of nanoscale communication networks will revolutionize health care. New modes of human-computer interaction will provide the basis for seamless integration of computing and communication technologies with our natural cognitive capacities.

We won't have to wait 2, 5, or 10 years for these technologies to provide immediate value. We are already seeing early advances in personalized medicine in terms of better retinal and cochlear implants, metered drug delivery, precisely targeted cancer treatment, and deep-brain-stimulation implants to relieve chronic depression, essential tremor and Parkinson's disease. More cumbersome instruments like MRI are being used to provide data on our emotional and cognitive states that can be used to train inexpensive, wearable devices that respond to our moods and preferences. These personal assistants will transform the entertainment business and early devices already seeing adoption in the gaming industry. New techniques in synthetic neurobiology show promise in the optogenetic control of the thalamus to interrupt seizures due to cortical injury, and every advance in the technology of efficient hybrid photovoltaics, genetically engineered biofuels and lab-on-a-chip microassays helps to move us closer to the goal of being able to monitor the brain at unprecedented scale and precision. Smart money will be watching for opportunties in all of these technology areas.

2.3.1 Introduction

Recent announcements of funding for two major initiatives [9, 121] in brain science have raised expectations for accelerated progress in the field of neuroscience and related areas of medicine. Both initiatives are depending on the development of new technologies and scientific instruments to realize their ambitious goals. Existing, well-established technologies will initially serve to propel the science forward, and these incumbent technologies will no doubt evolve to address new questions and offer new capabilities. However, we believe that current technologies will fall short in scaling to provide an appropriately wide range of co-registered assays across the whole brains of awake behaving animals and, in particular, human subjects[1]. We anticipate the need for new instruments that record diverse indicators of neural structure and activity[2] with substantially

[1]The EU funded Human Brain Project [121] (HBP) and the US funded Brain Research through Advancing Innovative Neurotechnologies (BRAIN) Initiative both motivate their research programs in terms of the potential benefit to human health and welfare. That said, there will continue to be a great deal of valuable research on alternative animal models, including primates, rodents and fish, as well as simpler organisms such as flies and other invertebrates. These models offer a variety of experimental options depending on the organism, including cloning, genetic engineering, and the application of alternative recording technologies, e.g., embryonic zebrafish are essentially transparent allowing neural imaging opportunities impossible in other organisms [192].

[2]The term "activity" is often used to describe the goals of the BRAIN (Brain Research through Advancing Innovative Neurotechnologies) Initiative. Indeed, the originally proposed name for the effort was the Brain Activity Map Project [9]. Unfortunately, the term is ambiguous even among neuroscientists and can be used refer to very different sorts of brain-related activity, including action potentials,

greater spatial and temporal resolution while providing rich contextual information of the sort required to relate these indicators to behavior and clinically relevant outcomes. In particular, this rich contextual information will be critical in analyzing recordings from multiple subjects or the same subject at different times and under varying conditions. There are precedents [408, 167] demonstrating how such scaling might be accomplished by taking advantage of the accelerating returns from computers, robotics and miniaturization that have made possible cell phones and the web. In addition, given their inherent potential for scaling, the nascent fields of synthetic biology and nanotechnology offer considerable promise but also pose formidable engineering challenges that will likely delay their availability as practical instruments for experimentalists.

If successful these instruments will produce a veritable tsunami of data well beyond the capacity of any existing computational infrastructure to cope; we will be buried in riches with no way to realize their value; even current recording technologies seriously tax existing industrial-scale computing infrastructure. The hope is that the accelerating returns from Moore's law and related trends in computing and storage technology will enable us to keep pace with the new technologies if we expend effort in developing and integrating appropriate data acquisition and analysis technologies throughout the scientific enterprise. The key is to automate as much as possible. One lesson from web-scale computing is that, if you aren't willing to automate an information processing task, then don't bother trying to carry it out by "hand", because soon enough someone else will have figured out how to automate it and they will run rings around you. The application of high-throughput methods adapted from genomics research coupled with new animal models and transgenic strains for drug screening is a good example of acceleration that leverages the scaling opportunities inherent in robotics, information processing technology and molecular biology.

There is no one technology that we can count on to make progress over the next decade. Practically speaking, the scientific community will have to move forward on a portfolio of promising technologies to ensure steady progress on short- and medium-term objectives while laying the technological foundations for tackling more ambitious goals. It should be feasible to engage the capital markets to accelerate progress and underwrite some of the development costs as the the technologies that drive the science will have important practical applications in communications and computing, medicine, and entertainment to name just a few of the relevant areas. The primary objective of this technical report is to populate the portfolio of promising technologies and provide scientists, entrepreneurs and investors with a basic understanding of the challenges and opportunities in the field as they are likely to evolve over the next decade.

We begin by making a distinction between recording and reporting. Conceptually every technology we discuss will consist of a *recorder* and a *reporter* component:

- recorder — think of a recording device that when you hit the record button converts energy from the microphone (sensor) into a semi-permanent record (physical encoding) of the signal for subsequent reuse; the notion of recorder combines the functions of sensing the target signal and encoding it a form suitable for transfer.

metabolic processes, e.g. mitocondrial efficiency, gene expression and diffuse neuromodulation, e.g. dopamine release in the *substantia nigra*, etc.

- reporter — think of a news reporter who seldom observes the actual events but rather collects first-hand accounts, writes her interpretation of events and posts them to her editor in a remote office; the notion of reporter combines the functions of transcoding, perhaps compressing and transmitting recorded information.

Recording technologies sample chemical, electrical, magnetic, mechanical and optical signals to record voltages, proteins, tissue densities, molecular concentrations, expression levels, etc.

There are many existing recording technologies we discuss in this report, but we focus our attention primarily on reporting as we view this as one of the primary bottlenecks limiting our ability to accelerate our understanding of the brain.

We divide recording technologies into two broad classes depending on whether the coding and signal-transmission components are located externally, outside of the target tissue, or internally, within the target tissue and typically co-located with the reporting components:

- external — an external energy source radiating in the electromagnetic or acoustic spectrum is used to illuminate reporter targets within the tissue and the reflected energy is collected and analyzed to read out measurements.

- internal — reporters are coupled to local transducers that pre-process and package measurements in a format suitable for transmission over a communication channel employed to transmit the coded signals to an external receiver.

Most external technologies involve some form of imaging broadly construed. In addition to those technologies explicitly described as "imaging", e.g., magnetic resonance imaging (MRI) and ultrasound imaging, we also include many of the technologies that incorporate the suffixes "scopy" or "graphy" in their name, e.g., scanning electron microscopy (SEM), near-infrared spectroscopy (NIRS), photoacoustic tomography and electroencephalography (EEG). External technologies have the advantage that most of the energy required to make the measurements, process the information and extract the data from the tissue can be supplied from sources outside of the tissue. The most common disadvantages of these technologies concern limitations in penetration depth due to scattering and absorption of energy within the tissue.

In contrast, internal technologies require a source of energy to power the implanted devices and a method of safely dissipating of the resulting waste heat. In addition, internal technologies generally require additional computational machinery to process, package and route raw measurements. In some cases, there are biomolecular computing solutions that are energy efficient and compatible with surrounding tissue. However, for many relatively simple computations we take for granted in conventional computing hardware, the biological options are too slow or error prone. Nanotechnologies based on either semiconductor technology or silicon-and-synthetic-biology hybrids hold promise if we can continue to drive down size and power and overcome problems with toxicity and potential interference with normal cell function.

2.3.2 Evolving Imaging Technologies

Many of the most powerful recording and reporting technologies currently available can be characterized as imaging broadly construed in which some portion of the acoustic or electromagnetic spectrum is used to illuminate a target tissue and the returning signal is analyzed to extract useful information about structure, e.g., cytoarchitectural details, and function, e.g., membrane potentials and spike-timing data. These relatively mature technologies largely finesse the problems relating to powering reporting devices and carrying out the required computations involved in signal processing, compression and transmission by performing all these functions externally to the imaged tissue.

These technologies are of particular interest since they are in daily use in hospitals throughout the world and are approved for human studies, unlike many of the other technologies we will discuss that are either not so approved or approved only in highly restrictive cases in which the patient has no other recourse for treatment or diagnosis. Both of the brain initiatives mentioned as depending on the accelerated development of the technologies surveyed in this report invoke understanding the human brain and relieving human suffering as motivation for their public funding. The imaging technologies discussed in this section are among those most likely to produce such outcomes in the near term.

With few exceptions, the technologies discussed in this section are relatively mature, were invented and refined largely independently of their applications in neuroscience, have promising directions for improved performance and extended capability, and are well funded by the private sector due to their use in the health care industry, industrial materials science and chip manufacturing. The physics of nuclear magnetic resonance (NMR) has given rise to a family of technologies that are familiar in modern medical practice but that are also commonly employed in scientific instruments used in many other fields. They include magnetic resonance imaging (MRI), functional MRI (fMRI), and diffusion tensor (functional) MRI (DTI).

fMRI works by measuring local changes in hemodynamic response (blood flow) that are roughly correlated with neural activity. The most commonly measured signal — blood-oxygen-level-dependent (BOLD) contrast [312] — serves as a rough proxy for neural activity averaged over rather large local populations of neurons and thus offers only limited spatial and temporal resolution. Studies involving fMRI offer some of the most powerful insights into high-level human cognition and its pathologies to date. Using fMRI scientists are able to decode patterns of activity involving cognitive functions identified with anatomically distinct areas of the brain such as auditory and visual cortex, and even implement a rudimentary sort of mind reading [289, 211]. State of the art fMRI resolution is on the order of 1 second temporal and 5 mm spatial.

The path from blood oxygen level to neural activity is anything but direct. Glucose plus oxygen yields energy in the form of ATP and waste products in the form of carbon dioxide and water[3]. This reaction takes place in mitochondria many of which are located in the synaptic knobs at the end of axons where most of the brain's metabolic budget is spent. We pay a high metabolic cost for processing information; it requires on the order of 10^4 ATP molecules to transmit a single bit at a chemical synapse [241]. Specifically, most of the ATP

[3]Oxygen is required in *glycolis* in which a glucose molecule is broken down into two three-carbon pyruvate molecules yielding two ATP molecules in the process.

consumed in the brain is spent pumping ions across cell membranes to maintain and restore resting potential following an action potential [161]. Despite its undisputed scientific and diagnostic merit, hemodynamic response and the BOLD contrast signal in particular are difficult to measure and complicated to interpret as an indicator for neural activity [245, 63, 265]. Better statistical tests to overcome measurement noise and alternative contrast agents are being sought after as alternative indicators for neural activity. Calcium-sensitive agents make MRI more sensitive to calcium concentrations and calcium ions play an important role as messengers for cellular signalling pathways in active neurons. Calcium indicators being developed for functional MRI may open up new opportunities for functional molecular imaging of biological signaling networks in awake, behaving organisms [16].

DTI estimates the rate of water diffusion at each location (voxel) in the imaged tissue. This enables us to identify white matter which consists mostly of glial cells and myelinated axons and comprises the major information pathways connecting different parts of the brain. DTI is particularly useful in a clinical setting for diagnosing stroke and white-matter pathologies. With a spatial resolution on the order of 100 μm it has provided some of the most stunning macroscale images to date for the Human Connectome Project (HCP) [394, 384]. HCP is tasked with understanding the micro- and macro-scale connections in the human brain. The general area of study is generally referred to as *connectomics*. We'll return to discuss microscale connectivity — connections between individual neurons — in Section 2.3.4.

There are a number of other imaging technologies that share the property they operate outside of the target tissue relying entirely on external sources of power. These include specialised conventional optical microscopes capable of μm resolution, and variants of confocal microscopy that achieve resolution below the diffraction limit to achieve resolution approaching 0.1 μm using laser light sources and super-resolution techniques [183, 373]. We will pay particular attention in the remainder of this report to the use of devices capable of sub micron resolution that are typically used in conjunction with various dyes and contrast agents to resolve the details of neurons including their dendritic and axonal processes and to identify the presence of proteins and other molecules of interest.

Scanning electron microscopes (SEM) can resolve details as small as a few nanometers, and, while not able to image individual atoms as are transmission electron microscopes, they are able to image relatively large areas of tissue which makes them the tool of choice for tracing neural circuits in chemically stabilized (fixed) tissue samples. Unlike MRI and ultrasound these technologies are primarily of use with fixed tissue, *in vivo* experiments or animal studies requiring invasive procedures not likely to ever be approved for human subjects. Two-photon excitation microscopy [176] is a fluorescence imaging technique that allows tissue penetration up to about one millimeter in depth and is particularly useful in circuit tracing and calcium imaging [154] in living tissue and genomic, transcriptomic, and proteomic maps of fixed neural tissue.

There are other imaging technologies relying on light or more exotic physics that deserve brief mention. Near-infrared spectroscopy (NIRS) is of interest as a non-invasive technology that can be used to detect a signal similar to BOLD but without an expensive and cumbersome MRI magnet. NIRS provides limited

spatial resolution and depth of field but is already proving useful as an inexpensive sensor in building brain-computer interfaces (BCI) for gaming headsets. Positron emission tomography (PET) provides some of the same capabilities as MRI but has received less attention in part due to its associated radiation risk and competing technologies maturing to subsume its most clinically relevant capabilities.

There is also a set of technologies that attempt to directly sense electrical signals resulting from neural activity without penetrating brain tissue and causing cellular damage. Electroencephalography (EEG) is perhaps best known of these technologies. EEG measurements can be recorded from awake behaving humans with no surgical intervention, but because the electrodes are placed on the skin covering the skull it offers a spatial resolution on the order of 5cm, and, while the temporal resolution can be as high 1KHz, measurements average over large populations of neurons. Electrocorticography (ECoG) allow us to measure local field potentials (LFPs) with temporal resolution of approximately 5 ms and a spatial resolution of 1 cm but involves a craniotomy as it requires a grid of electrodes placed directly on the exposed surface of the brain. Advanced ECoG methods have successfully decoded speech from patterns of activity recorder over auditory cortex [322].

Magnetoencephalography (MEG) — like EEG and ECoG — measures the net effect of currents flowing through the dendrites of neurons during synaptic transmission — not, it is worth noting, action potentials. The synchronized currents of neurons generate weak magnetic fields just barely detectable above the background levels even with extensive shielding and superconducting magnetometers. On the order of 50,000 active neurons are required to generate a detectable signal. MEG has the temporal resolution of EEG and the spatial resolution of ECoG, however unlike ECoG MEG does not require a craniotomy.

Light in the visible range of the electromagnetic spectrum cannot penetrate deeply into tissue. Clearing reagents like CLARITY [87] and Clear$^{\mathrm{T}}$ [237] can be used to prepare tissue in such a way that structures that would normally absorb or scatter light such as the bilipid layers that comprise cell membranes are rendered transparent, but this approach doesn't apply to living tissue. At radio frequencies below 4 MHz the body is essentially transparent to the energy which makes it a candidate for transmitting data but not imaging. Light in the near-infrared range with a wavelength of about 800 nm to 2500 nm can penetrate tissue to several centimeters and is used in both imaging and stimulating neural tissue. However, the electromagnetic spectrum is not the only alternative we have for non-invasively probing the brain.

Ultrasound pressure waves are routinely used in diagnostic medical imaging and easily penetrate tissue to provide real-time images of the cardiovascular system. The ultrasonic frequencies used for diagnostic imaging range typically between 2 MHz and 20 MHz. Spatial resolution on the order of 1 μm is possible with frequencies in the 1-2 GHz range, but attenuation in soft tissue increases as a function of the frequency thereby reducing penetration depth[4]. This loss of penetration depth can be compensated somewhat by increasing the signal intensity while limiting exposure. For exposures longer than ten seconds, intensity levels less than 0.1

[4]Theoretically, attenuation increases with the square of the frequency, but linear increases have been reported for several biological tissues [182]. Unfortunately, the empirically-derived attenuation coefficient for neural tissue is 200 times that of water — 0.435 versus 0.002 (db cm^{-1} MHz^{-1}) [182] — despite the two materials having similar density, 1.0 versus 1.02 (g/cm^3), speed of sound, 1480 versus 1550 (m/sec) and acoustic impedance, 1.48 versus 1.60 ([kg/(sec m^2)] $\times$ 10^6).

(W/cm^2) are generally deemed safe for diagnostic imaging. For a 70 MHz signal traveling through water, the attenuation coefficient is 10 (dB/cm) and, given an intensity of 0.1 (W/cm^2), the maximum effective depth would be in the range 1.5-2 cm.[5].

Ignoring temperature and barometric pressure, ultrasound travels about five times faster in water (about 1500 m/s) than it does in air (about 300 m/s). Most ultrasound technologies exploit the piezoelectric effect for acoustic signal transmission, reception or both. When materials like ceramics, bone and DNA are mechanically stressed, they accumulate an electrical charge. Conversely, by applying an external electric field to such materials, we can induce a change in their static dimensions. This inverse piezoelectric effect is used in the production of ultrasonic sound waves.

Focused ultrasound (FUS) technologies developed for medical applications employ a phased-array of piezoelectric transducers to produce multiple pressure waves whose phase is adjusted by introducing delays in the electrical pulses that generate the pressure waves. By coordinating these delays, the focal point — point of highest pressure and thus highest temperature in the tissue — can be precisely controlled to avoid cell damage. FUS can be used for deep brain stimulation and has demonstrated promise in clinical trials on patients suffering from essential tremor. FUS has also been used to alter the permeability of the blood-brain barrier to allow the controlled diffusion of drugs or other nanoparticles across the blood-brain barrier [450]. A hybrid near-infrared-plus-ultrasound-imaging technology called *photoacoustic spectroscopy* has been shown effective in monitoring focused-ultrasound-induced blood-brain barrier opening in a rat model *in vivo*.

High-intensity focused ultrasound (HIFU) can be used to destroy a tumor in the breast, brain or other tissue without cutting into the surrounding tissue. In the case of the brain, the cranium poses a challenge due to its variable thickness, but this can be overcome either by performing a craniotomy or by using a CT scan to construct a 3-D model of the skull and then generating a protocol based on this model that adjusts the delays to correct for aberrations in signal propagation due to the changes in thickness of this particular skull [431]. HIFU offers an alternative to Gamma-knife surgery without the attendant radiation risk though there are some drawbacks due to the fact that ultrasound waves, unlike ionizing radiation, can be deflected and scattered.

Ultrasound technologies like transcranial doppler (TCD) imaging have been used to measure the velocity of blood flow through the brain's blood vessels, and provide fast, inexpensive diagnostics for brain injuries [28]. TCD offers temporal resolution comparable to other neuroimaging techniques, but spatial resolution is relatively poor at the frequencies typically used in diagnostic imaging. High-frequency acoustic microscopes do exist, however, and are routinely employed in clinical settings, most notably in producing *sonograms* used in ophthalmology for treating glaucoma [388]. For the time being, it seems the most likely applications of ultrasound in experimental neuroscience will involve diffuse stimulation using nanobubble contrast agents [105] and manipulation of the blood-brain-barrier for introducing engineered biomolecules

[5]The formula for the attenuation coefficient in water as a function of frequency is $\alpha = 2.17 \times 10^{-15} \times f^2$ (dB/cm) where f is the frequency [295]. In order to achieve a spatial resolution of 15 μm, we would need a frequency of 100 MHz and given that the corresponding α is approximately 20 (dB/cm) it would be difficult if not impossible to safely penetrate to the maximum depth required to image an entire human brain.

and nanoparticles into the brain — see Appendix 2.3.11.

The technologies discussed in this section are relatively mature. In some cases, they have run up against fundamental physical limitations, and improvements in resolution and accuracy are likely to come from incremental engineering. That said they are also the incumbents in the extant technology race; their technology has set a high bar, and even incremental improvements will likely prove sufficient to maintain their dominance and market share. In many cases, they also have the advantage that the medical profession is conservative and reluctant to abandon technology in which they have invested time learning how to master. To break into these established markets, new companies will have to demonstrate substantially new capabilities to attract funding and successfully launch products.

Attractive capabilities that might serve as game changers include technologies that combine recording and stimulating, e.g., thus enabling the use of feedback for direct control of treatment, multi-function technologies that monitor several indicators at once, e.g., recording co-located electrical and proteomic activity, lighter, smaller, less intrusive technologies, e.g., wearable or implantable devices to support remote monitoring and improve patient compliance, and technologies that significantly expedite existing protocols or eliminate them entirely, e.g., portable, affordable alternatives to diagnostic MEG, MRI and PET for small family medical practices.

Speed definitely matters in both clinical and scientific studies. Multi-plane, parallel-imaging NMR technologies[6] have accelerated scanning four-fold, and there are likely greater gains to be had as other parts of the processing pipeline catch up. With the advent of portable devices for bedside ultrasound scanners and a move by manufacturers to support beam-forming in software researchers have been able to improve throughput thirty-fold with no reduction in resolution [263]. The latest SEM technology promises high-throughput, large-area imaging using a multi-beam scanning. The electron source is split into multiple beams — as many as sixty — and all of the beams are scanned and recorded from in parallel resulting in a sixty-fold speedup in acquisition.

While some improvements among the incumbent technologies will require fundamental advances in science, improvements requiring computation can immediately take advantage of the accelerating returns from advances in computer design and chip fabrication. For example, new signal-processing algorithms running on faster hardware can increase accuracy with no loss in throughput by sampling more frequently, sampling more intelligently or using existing sampling strategies but spending more time analyzing the samples. MRI, ultrasound imaging and scanning electron microscopy are all poised to make substantial improvements in throughput by exploiting computation to accelerate the acquisition and analysis of data.

2.3.3 Macroscale Reporting Devices

In this section we continue our discussion of relatively mature technologies by examining tools for reporting on the microscale properties of individual cells using technologies whose components are implemented at the

[6]Parallel acquisition techniques combine the signals of several coil elements in a phased array to reconstruct the image, the chief objective being either to improve the signal-to-noise ratio or to accelerate acquisition and reduce scan time.

macroscale. In Sections 2.3.5 and 2.3.6 we will return to the problem of resolving details at the microscale but this time employing microscale components based on, respectively, re-purposed biological components and hybrid technologies that combine biological parts with inorganic devices including novel applications of semiconductor and integrated-circuit technology.

The classic single-probe electrode used in electrophysiology consists of a solid conductor in the form of a needle that can be inserted into the tissue to record the local field potentials resulting from the activation of neurons in close proximity to the uninsulated tip of the electrode. In Section 2.3.4 we review current methods of automating the insertion and manipulation of such devices thereby eliminating the most time- and labor-intensive part of their application in the lab.

While still used in practice many scientists now employ multi-electrode arrays consisting of hundreds of electrodes arranged in a grid in order to better resolve the activity of individual neurons. These arrays can be implanted in living tissue for *in vivo* experiments or cells can be cultured on arrays for *in vitro* studies. Arrays constructed from stainless steel or tungsten are now being replaced by silicon-based devices that can benefit from the same technologies used in chip manufacture, and technologies originally intended for the lab are being refined for use as permanent implants in humans spurring innovation in the design of biocompatible devices.

Tightly grouped bundles of very small electrodes are often used to enable more accurate local readings. Four such electrodes in an arrangement called a *tetrode* provide four spatially-distributed channels that can be used to better separate the signals originating from different neurons. The basic idea can extended and *polytrodes* consisting of 54-channel high-density silicon electrode arrays have been used to make simultaneous recordings from more than 100 well-isolated neurons [40]. Flat arrays called *microstrips* with as many as 512 electrodes and 5 μm spacing have been used to record from and stimulate cells in the retina and visual cortex [254].

The recording tips of the electrodes in the multi-electrode arrays mentioned above are typically arranged to lie on the same plane and so these arrays are essentially 2-D probes. To enable an additional degree of freedom in recording, these probes can be advanced or retracted in small steps to sample from a 3-D volume, but this implies that at every point in time the samples are all drawn from a planar region. It is worth noting that while the tools of electrophysiology are used primarily for sampling extracellular voltages, electrophysiology can also measure internal voltages of a single neurons. In studying hippocampal place cell activity, sub-threshold dynamics seem to play an important role and there are currently limited ways to get a similar signal with the other technologies [162].

Single probes with multiple recording sites along their length have been developed with as many as 64 channels [110]. These probes are particularly revealing when inserted into cortical tissue either vertically or horizontally relative to the cortical surface to record from different types of neurons located in multiple layers or from multiple neurons of the same type of neuron located within the same layer. If you ask an experimental neuroscientist interested in early vision if it would help to record more densely from visual cortex, the answer would likely be yes since it is a common and puzzling experience having recording from a

neuron that appears to strongly respond to one stimuli, only to find when you move your probe scant microns you encounter a neuron that responds strongly to a completely different stimuli.

There is also work developing true 3-D arrays using flexible materials for use in chronic implants [78]. Optogenetics has opened the door to implants that can exert exquisite control over individual neurons with optical-fibers being the method of choice for delivering light to exert such control. In principle the same wave guides used to deliver light to activate neurons can be multiplexed to receive light from biomolecules designed to record from neurons. Optical wave guides that can be used for delivering light to excite or inhibit individual neurons in a 3-D volume have been demonstrated [482] and since these probes were fabricated using CMOS technology the expectation is that these technology will scale to thousands of targets with the 3-D volume.

A *fluorophore* is a fluorescent compound that re-emits photons upon excitation and can be covalently bonded to a macromolecule to serve as a dye, tag or marker. Fluorophores can be employed as microscale recorders capable of resolving nanoscale details. The method of *calcium imaging* is perhaps the most successful such application of this idea to recording from many neurons simultaneously [155]. The synaptic transmission of action potentials is controlled in part by an influx of calcium into the synaptic terminal of transmitting neuron's axon[7]. The distribution of calcium in synapses can be used as a proxy for neural activity. Detection is accomplished using genetically encoded calcium indicators (GECI) that respond to the binding of $Ca2^{+}$ ions by changing their fluorescence properties [192].

Imaging is typically accomplished using one- or two-photon microscopy but it was recently shown to be possible to image the entire brain of a larval zebrafish at 0.8 Hz and single-cell resolution using laser-scanning light-sheet microscopy [5], and, despite calcium being a lagging indicator, it is possible to employ calcium imaging data to reconstruct spike trains using Monte Carlo sampling and super-resolution techniques [441]. Miniaturized fluorescence microscopes offering ˜0.5 mm^2 field of view and ˜2.5 micron lateral resolution enable researchers to record from awake, behaving mice [137].

Patch clamping is a bench technique for recording from single or multiple ion channels in individual neurons. It is the most precise method available to the neuroscientist for recording electrical activity in neurons. Until recently it required a highly skilled technician to carry out the procedure. The method has now been automated [222] opening up the possibility of highly-parallel, robotically-controlled experiments. It is possible to build nanoscale recorders that indicate whether a specific ion channel is open or closed, and so it is natural to ask if one could not simply record from *all* the ion channels on an axon and use this information to reconstruct the propagation of action potentials. While conceivable in principle, it is not likely to prove practical any time soon due to the simple fact that there on the order of 1,000 sodium pumps (voltage-gated ion channels) per μm^2 of axonal membrane surface or about a million sodium pumps for a small neuron.

It is possible to directly measure membrane potentials optically by using a combination of two-photon microscopy and genetically-encoded voltage indicators (GEVI) derived from a combination of a voltage-sensing domain similar to that found in voltage-gated ion channels and fluorescent proteins. GEVI proteins

[7]There is also post-synaptic (somal) calcium influx during depolarization, and for this reason the method of calcium imaging can be used to analyze the antecedents to action potentials in firing neurons not just the consequences manifest in their axonal processes.

can be targeted to specific cells and there expressed and integrated into the cell membrane [370]. However, instead of opening an ion channel, the voltage-sensing domain serves to alter the conformation of the fluorescent proteins so as to signal changes in membrane potential [327]. Researchers have demonstrated that it is possible to use such a method to reliably detect single action potentials in mammalian neurons [230]. More recently, both subthreshold events and action potentials have been measured in genetically-targeted neurons in the intact Drosophila fruit fly brain [72]. While promising, the technology has yet to seriously challenge calcium imaging due in part to problems with signal-strength and temporal-response limitations.

There are several relatively static pieces of information that neuroscientists would like to have for their experimental subjects: the set of all RNA molecules transcribed in each cell type — referred to as the *transcriptome*, the set of all proteins expressed by each cell type — called the *proteome*, and a complete map of the connections between cells along with some measure of the strength of those connections referred to as the *connectome*[8]. In terms of basic genomic information, we might sequence an instance of each cell type or multiple instances of the same cell type drawn from different locations in order to search for epigenetic differences. However, full sequencing is likely unnecessary given that we know what to look for and so, for practical purposes, it should suffice to apply a method such as *in situ* hybridization or immunoflorescence to obtain genetic signatures for cells or create a map relating genetic differences to locations corresponding to the coordinates of small volumes of tissue.

One point to emphasize is the importance of maps and why they are so useful to scientists. In particular, we envision multiple maps: genomic, connectomic, transcriptomic, and proteomic maps along with sundry other maps registering activity and metabolic markers. These maps will be significantly more useful if different maps from the same subject can be registered with one another so we can relate their information spatially — and temporally in the case of maps with a temporal dimension. Equally important is the ability to relate maps from different subjects of the same species, as in the case of comparing healthy and diseased brains. As simple as this sounds, it is an enormously complex problem. That the Allen Institute was able co-register the maps of almost identical cloned mice in building mouse-brain atlas [408] is a significant accomplishment and preview of some of the difficulties likely to be encountered in building a human-brain atlas. Ideally we would like a standardized method for moving between organisms; such methods exist but terminology differences between humans and model organisms complicate the mapping. A controlled ontology analogous to the NIH-maintained Medical Subject Headings (MeSH) for neuron cell types, neuroanatomical landmarks, etc. would enable more fluid transitions between different experimental results.

Though there are many variations the most common approach to generating a connectome involves stabilizing and staining the tissue sample, slicing it into thin sections, scanning each slice with an electron microscope, and then applying some method to segment the cell bodies and identify connections between cells [54, 284]. This last stage is the most time consuming by several orders of magnitude as it typically involves the intervention of a human expert. While automating cell-body segmentation is still an open problem,

[8] If exhaustive, the connectome would have to include axodendritic, axosomatic and dendrodentritic connections along with a classification of the pre- and post-synaptic cell types, relevant neurotransmitters, and information on whether the connections are excitatory or inhibitory.

most computer vision experts believe that it will be solved by some combination of better algorithms, better stains and better imaging hardware [193]. Refinements of tissue-preparation technologies such as CLARITY [87] and Clear^{T} [237] — see Section 2.3.2 — may eventually prove useful in segmenting cell bodies. We will return briefly to the problem automating the production of connectomes in Section 2.3.4.

Array tomography is a method for extracting a proteomic map which is superficially similar to the method described above for producing connectomes in all but the last step. Instead of applying a computer vision algorithm to segment each SEM image, you apply the method of immunofluorescence to tag proteins with fluorescent dyes so they can be identified with a scanning electron microscope [283]. Alternative methods of tagging such as fluorescent *in situ* sequencing (FISSEQ) on polymerase colonies [290] may offer a way around current limitations in the number of tags — and hence proteins — you can distinguish between in a single tagging-and-imaging pass over a given volume.

A volumetric transcriptomic map identifies for each 3-D volume in a tissue sample the set of all RNA molecules produced in the population of cells located within the volume. Ideally such a map would have a temporal extent given that the type and number of transcribed genes will vary over time, at the very least over the development of the organism [207]. The simplest approach to constructing a static volumetric transcriptomic map is to slice the sample into small volumes, separate out the RNA and sequence using RNA microarrays, FISSEQ or some other high-throughput sequencing method.

In this section, we looked at reporting technologies consisting of macroscale components which, while capable of reporting out the information provided by microscale recorders, are volumetrically limited by absorption and scattering. The exception being static maps of tissue samples that are dissembled for analysis assuming that the organism can be sacrificed. By employing a fiber-optically coupled microendoscope [27] it is possible to image cells deep within the brains of live animals over extended periods, but with reduced field of view and the invasive introduction of the endoscopic device. Dynamic maps that record neural activity in awake behaving humans are limited in scale by existing reporter technology. In Sections 2.3.5 and 2.3.6, we will explore technologies for implanting microscale reporter co-located with their microscale recorder counterparts to enable scaling activity maps to even larger spatially distributed ensembles of neurons.

2.3.4 Automating Systems Neuroscience

The notion of testable hypothesis and the quest for simple, elegant explanatory theories is at the very foundation of science. The idea of automating the process of hypothesis generation, experimental design and data analysis, while once considered heretical, is now gaining favor in many disciplines, particularly those faced with explaining complex phenomena. The search for general principles — the holy grail of modern science — sounds so reasonable until you ask scientists what would qualify for such a principle, and here you are likely to get very different answers depending on whom you ask. We would like the world to be simple enough accommodate such theories, but there is no reason to expect nature will cooperate. Perhaps in studying complex systems like the brain, we'll have to settle for a different sort of comprehension that speaks to the emergent properties and probabilistic interactions among simpler components and mathematical characterizations of

their equilibrium states [9, 341, 181].

The term "emergent" is often used derisively to imply inexplicable, opaque or even mystical. However, there is reason to believe at least some of the processes governing the behavior of neural ensembles are best understood in such terms [405, 64]. This is not a call for abandoning existing theories or the search for general principles, but rather a suggestion that we consider new criteria for what constitutes an adequate explanation, entertain new partnerships with computational scientists and where possible automate aspects of the search for knowledge, and explore new classes of theory that demand data and computation to discover and evaluate. Computation- and data-driven approaches don't necessarily imply recording from millions of neurons, though techniques from data mining and machine learning may be the best tools for making sense of such data. Indeed, there already exist powerful tools in the current repertoire — see Sections 2.3.2 and 2.3.3 — that can be applied to produce data of sufficient quantity and quality to explore interesting such phenomena [44, 71, 394]. Distributed and sparse coding models of visual memory were initially motivated by computational and statistical arguments [26, 316] and subsequent work developing probabilistic models crucially depend on analyses that would not have been possible without modern computational tools and resources [381, 71, 358].

As our understanding of smaller circuits and diverse molecular, electrical and genetic pathways improves we can expect to see increased reliance on high-fidelity simulations and data-driven model selection[9] [403, 417]. These approaches will enjoy accelerated returns from advances in computer technology allowing scientists to explore and model larger ensembles. In the past, mathematicians, statisticians and computational scientists interested in analyzing neural data often found themselves isolated at the far end of a pipeline, shut out of the preliminary discussions involved in designing experiments and vetting theories, and privy to such earlier decisions only through the spare, stylized format of academic publication. This sort of specialization made it difficult to approach the problem of model selection in a systematic end-to-end fashion with opportunities to adjust each step of the process from stipulating which measurements are taken and how the data is annotated, curated and stored electronically to defining which hypotheses and classes of models are appropriate in explaining the data. Today computational scientists are integral members of multidisciplinary teams.

The idea of including computational scientists early in designing experiments, vetting theories and creating models and frameworks for understanding is not new. Neural simulators have been around as long as computers[10], and modern high-fidelity, Monte Carlo simulations [213] are capable of modeling the diffusion and chemical reactions of molecules in 3-D reconstructions of neural tissue and are used for a wide range of *in silico* experiments, e.g., providing evidence for ectopic neurotransmission in synapses [94] and accurate

[9] In statistics and machine learning, *model selection* is the problem of picking from among a set of mathematical models all of which purport to describe the same data set. The task can also involve the design of experiments to ensure that the data collected is well-suited to the problem of model selection.

[10] In developing the Hodgkin-Huxley model of action potentials [180], Huxley carried out extensive calculations on a manually-cranked calculator to make predictions about how action potentials would change as a function of the concentration of extracellular sodium. His predictions were later confirmed by Hodgkin's experiments with giant squid axons [179]. Around the same time, a model of synaptic plasticity developed by Donald Hebb [173] was simulated on an early digital computer at MIT by Farley and Clark in 1954 [127].

simulations of 3-D reconstructions of synapses yielding new insights into synaptic variability [402].

While several large-scale simulations [12, 191, 271] have received attention in recent years, skeptics believe that such efforts are premature given the present state of our knowledge [442]. Whether true or not, many researchers characterise their work as trying to understand how neural circuits give rise to behavior. But the gap between circuits and behavior is wide, and an intermediate level of understanding might be achieved by characterizing the neural computations which occur in populations of neurons [73]. In the same way that knowing the primitive operators in a programming language is essential to understanding a program written in that language, so too the computations we use to characterize the function of smaller circuits might eventually provide us with a language in which to describe the behaviors supported by circuits comprised of larger ensembles of neurons.

In some cases, it is instructive to discover that one sort of information, say spiking behavior, can be recovered from another sort of information, say calcium influx, using an appropriate machine learning algorithm. Computational neuroscientists demonstrated that accurate spike-timing data could be recovered from the calcium imaging of neural populations opening the possibility of recording from neural circuits and inferring spikes — the gold standard for summarizing the electrical behavior of individual neurons — without patch clamping or inserting electrodes [441]. In developing the Allen Mouse Brain Atlas [408], the Allen team had to invent techniques for registering neural tissue samples from multiple cloned mice against standardized anatomical reference maps and, in the process, they developed powerful new tools for visualizing 3-D connectivity maps with detailed gene-expression overlays. Patterns apparent from gene expression maps are often as useful if not more so than canonical reference maps for identifying functional areas, and this observation may be of crucial importance when we attempt to build a human brain atlas.

It is worth mentioning that the Allen Mouse Brain Atlas required the collaboration of scientists from diverse fields including neuroanatomy, genomics and electron microscopy and would not have been possible without the involvement of mathematicians and computer scientists from the very start of the project. Given the trend, we expect more successful collaborations involving neuroscientists and biophysicists with a deep understanding and broad interpretation of computation as manifest in biological systems and computer scientists and electrical engineers with extensive knowledge and appreciation of biological organisms and the many chemical, electrical, mechanical and genetic pathways that link those organisms to the environments in which they evolved. And in, terms of leveraging tools, techniques and technologies that benefit from the accelerating returns of computation, we can also expect help from another unexpected quarter: industrial and consumer robotics.

The automation of serial sectioning, section handling and SEM imaging[11] for large-volume 3-D reconstruction and array tomography has enormously sped up the collection of data required for applications in connectomics and proteomics. Developments like the automatic tape-collecting ultra-microtome (ATUM) remove the requirement for a skilled human in the loop and dramatically reduce sorting errors and tissue-handling damage [371]. The automated 3-D histological analysis of neural tissue is still in the early stages of

[11] As discussed in Section 2.3.3, serial-section block-face scanning electron microscopy is one of the primary methods used in connectomics and proteomics for creating 3-D maps of stabilized, stained neural tissue [53].

development but a great deal of progress has been made [287] and we can expect rapid improvement in the next few years fueled by advances in machine learning [193] and high-performance computing.

Indeed, there are many repetitive tasks routines carried out in the lab that can be automated by machine learning and computer vision or accelerated by robotically controlled instruments. Even such delicate tasks as patch-clamp electrophysiology [222] and multi-electrode array insertion [481] can be performed by programmable robots. Increasingly the manufacturers of scientific instruments are replacing hardware components by software components so they can offer performance enhancements by simply upgrading the software or swapping out a circuit board and replacing it with one using the latest processor technology.

The ability to run hundreds of laboratory experiments in parallel with little or no human intervention is now possible with modular robotic components, standardized controllers and computer-vision-enabled monitoring and data collection. Of course, prior to automating a complicated endeavor like that addressed by the Encyclopedia of DNA Elements (ENCODE) Consortium[12], you first have to do a lot of exploratory work to figure out what's worth automating. That said, researchers should be aware of the potential advantages of automation tools and quick to exploit them when they identify an appropriate task. The short- and medium-term prospects for more parallelism, higher throughput, greater precision and enhanced flexibility is limited primarily by our imagination and willingness to invest in building and deploying the requisite systems.

2.3.5 Synthetic Neurobiology

In this section, we consider the methodology of co-opting existing biomolecules for purposes other than those for which they naturally evolved. This approach has the advantage that it often simplifies the problem of biocompatability. Moreover it allows us to take advantage of the enormous diversity of solutions to biologically relevant problems provided by natural selection [359]. Given that cellular function is conserved across a range of organisms from algae to mice, if an existing solution from the target organism is not found, a solution from an alternative organism is often compatible.

We've already seen several examples of biomolecules that play important roles in technologies relevant to scalable neuroscience. For instance, organic fluorophores are used in imaging technologies to stain tissues or as markers for active reagents, as in the case of antibodies used in immunoflorescence imaging. Biomolecules found in odd flora and fauna have found application in genomics and provide a dramatic example of how biology can enable exponential scaling.

A key step in DNA sequencing involves amplifying the DNA to create the many copies required in subsequent steps such gel electrophoresis[13]. Amplification requires multiple cycles in which a heat-stable DNA polymerase plays a critical role. An important breakthrough was the discovery of Taq polymerase isolated from *Thermus aquaticus* a species of bacterium that can tolerate high temperatures, but it required considerable additional effort before the method now called polymerase chain reaction (PCR) was refined

[12]The goal of ENCODE is to build a comprehensive parts list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active.

[13]It is worth noting that this amplification step will no longer be necessary in nanopore sequencing or other singlemolecule sequencing methods [434].

and became an indispensable tool in genomics [300].

The discovery of complex molecules called *channelrhodopsins* played a similar role in spurring the development of optogenetics [47]. These molecules function as light-gated ion channels and serve as sensory photoreceptors in unicellular green algae to control movements in response to light. We'll look at the development of optogenetics in more detail as a case study in how technologies in this field develop and a object lesson in the difficulty of predicting how long it takes to propel an idea from its initial conception to a useful technology.

Once you've identified a suitable molecule to perform your target function, you have to figure out how to introduce it into the cell. In some cases, it is possible to introduce the molecule directly into the intra- or extra-cellular fluid, but more often than not, it is necessary to enlist the cell's machinery for replication, transcription and translation to produce the necessary active molecules and control their application. This last typically involves the technology of recombinant DNA and synthetic biology.

There are a number of technologies for inserting genes into a host cell's DNA. In some cases, you can co-opt an existing pathway that will express and control the expression of a protein in exactly the right way. It is also possible to modify an existing pathway or create an entirely new pathway, but such modifications and additions are notoriously difficult to get right and can delay or derail development. Such problems are exacerbated in the case of introducing additional compounds into the mix.

Despite its challenges, this approach to building nanoscale recording and reporting devices is promising since the issues of biocompatibilty which plague more exotic approaches based on nanotechnology are much more readily solved, given the current state of the art. Moreover, these solutions utilize existing cellular sources of energy and biomolecular processes, many of which can be viewed as carrying out information-processing tasks, are remarkably energy efficient compared to semiconductor technology. To illustrate, we look at three examples:

The role of competition, rich collaboration, timing, serendipity and just plain luck make it difficult to predict when or even whether an idea, however compelling, will mature to point that it serves as a useful tool to enable new science. This is nowhere more apparent than in the case of optogenetics, one of the most powerful new technologies to emerge out of systems neuroscience in the last decade [469].

The basic molecules used in optogenetics — called *opsins* — have been studied since the 1970s. These complex proteins undergo conformational changes when illuminated that serve to control the transfer of specific ions across the membranes of cells in which they are expressed. Found in archaea, bacteria, fungi and algae these large-molecule proteins serve diverse photosynthetic and signaling purposes. While obvious in hindsight, at the time it was not obvious that these proteins could be re-purposed to control the firing of neurons.

Once the basic idea was formulated, the search was on for a molecule that could be expressed in mammals at levels high enough to mediate neural depolarization, was well tolerated at such levels, didn't require additional reagents to enable *in vivo* experiments, and recovered quickly enough to allow for precise control of the target neurons. There were also the technical problems of how to introduce the molecule into cells

and how to precisely deliver light to the target neurons that needed to be solved to provide a compelling demonstration.

Before the field stumbled upon channelrhodopsin there were a bunch of alternative technologies being considered including ones with longer chains of dependencies and more complicated component technologies involving both natural and synthetic options both biological and inorganic. This is often the natural chain of events leading to a simpler, more elegant solution matching the task to a specific solution in nature that solves the problem.

Once the idea was out, several labs came out with related technologies around the same time. There were also plenty of exciting innovations including a demonstration of light-activated neural silencing in mammals and refinements such as a channelrhodopsin that can be opened by a blue light pulse and closed by a green or yellow light pulse. Opsins probably could have been applied in neuroscience decades earlier [47], but excitement and competition often serve as the tinder required to ignite a flurry of rapid development [474].

Generating the connectome at the macroscale — tracing the major white-matter connections between functional areas — is reasonably tractable using diffusion tensor MRI. Generating the microscale connectome — mapping the connections between individual neurons — is much more challenging and it may be that approaches based on analyzing scanning electron micrographs are tackling a harder problem than is strictly necessary. Might there be a simpler way of creating a catalog of all the neurons in a tissue sample, their connections, cell type, connection strengths and 3-D coordinates?

Zador *et al* [471] have proposed the idea of "sequencing the connectome" in a paper of the same title emphasizing the opportunity to leverage one of the most important scalable technologies to emerge from the biological sciences. The basic concept is simple. The authors break down the problem into three components: (a) label each neuron with a unique DNA sequence or "barcode"', (b) propagate the barcodes from each source neuron to each synaptically-adjacent sink neuron — this results in each neuron collecting a "bag of barcodes", and (c) for each neuron combine its barcodes in source-sink pairs for subsequent high-throughput sequencing.

All three steps — excluding the sequencing — are to be carried in the live cells of the target tissue sample. The details are not nearly as simple and each step will likely require significant engineering effort. However, as in the case of optogenetics, there are a number of tools that might be applied directly or adapted to suit. There exist reasonable approaches for generating random barcodes from DNA templates that could be used as unique designators for individual neurons.

We also know how to propagate barcodes transynaptically using viruses that naturally spread from neuron to neuron through synapses. The operation of "concatenating" barcodes is carried out in any number of neurons. Sequencing and location tagging would be accomplished using the method described in Section 2.3.3 for generating a static volumetric transcriptomic map. Getting all these pieces to play together and not compromise the cell before all the connectomic data is collected might be the work of a summer or several years. It is difficult to predict the outcome, but several experts believe Zador's approach or something like it will work and the first demonstrations on simple organisms are one to two year out.

Calcium imaging offers our best near-term scalable method for recording from large ensembles of living neurons in mice or even simpler experimental animals. But what if we want to record from neurons deep within the brain in an awake behaving human? One proposal is to design a molecular machine to sense a convenient correlate of neural activity such as elevated calcium concentrations and write the resulting data to a molecular "ticker tape" [473, 228]. While somewhat more subtle than the previous proposal, the individual steps are conceptually straightforward.

Cellular transcoding processes such as DNA replication and messenger RNA transcription make occasional errors in the process of incorporating nucleotides into their respective products. The proposed method depends on harnessing an enzyme called DNA polymerase (DNAP) that plays a central role in DNA replication. Relying on the property that the rate of misincorporation is dependent on the concentration of positive ions or *cations*. They show that by modulating cation concentrations one can influence the misincorporation rate on a reference template in a reliable manner so that information can be encoded in the product of DNAP and then subsequently recovered by sequencing and comparing with the reference template.

There are a lot of moving parts in the ticker-tape technology. A full solution will require packaging all the necessary enzymes, nucleotide-recycling, and metabolic machinery necessary to initiate and sustain DNAP reactions outside of the nucleus. This idea may be in the early, wishful-thinking stage of exploring complex, multi-step solutions, hoping that in the meantime some molecular biologist will stumble across an existing ticker-tape-like solution just waiting for us to exploit[14]. It is very hard to make any reasonable predictions of how long it will take this technology to mature or if it will resemble anything like the current proposal if a solution is ever realized. Two to five years is optimistic and a final solution will likely bear little relationship to the current proposal.

It is no longer necessary to depend solely on what we discover in the natural world when searching for biomolecules for engineering purposes. Methods from synthetic biology like rational protein design and directed evolution [45] have demonstrated their effectiveness in synthesizing optimized calcium indicators for neural imaging [192] starting from natural molecules. Biology provides one dimension of exponential scaling and additional performance can be had by applying high-throughput screening methods and automating the process of cloning and testing candidates solutions.

There are also opportunities for accelerated returns from improved algorithms for simulating protein dynamics to implement fast, accurate, energy function used to distinguish optimal molecules from similar suboptimal ones. Advances in this area undermine arguments that quantum-dots and related engineered nanoscale components are crucial to progress because these technologies can be precisely tuned to our purposes. Especially when weighed against the challenges of overcoming the toxicity and high-energy cost of these otherwise desirable industrial technologies.

Biology doesn't offer any off-the-shelf solutions addressing the reporting problem aside from writing the recorded data to some stable molecular substrate like DNA and then flushing the encapsulated data out of

[14]Finding a DNAP sensitive to calcium or other biologically relevant ions is difficult, but, as in the case of channelrhodopsin and Taq polymerase, a novel organism's DNAP might come to the rescue, especially as it becomes clearer exactly what we're looking for.

the nervous system to be subsequently recovered from the organism's blood or urine[15] We aren't likely to find solutions to the problem of transmitting large amounts of data through centimeters of nervous tissue for the obvious reason that nature hasn't come across any compelling need to read out the state of large neural populations. To engineer scalable solutions to the reporting problem, we may have to look to advances in building nanoscale electronic circuits.

2.3.6 Nanotechnology

Nanotechnology, the manipulation of matter at the atomic scale to build molecular machines, is already contributing to brain science [8]. Much of the recent progress in developing probes for multi-cell recording depends on nanotechnology [482, 27, 110, 78]. From advances in chip fabrication for faster computers to better contrast agents for imaging and high-density, small-molecule assays for high-throughput pathogen testing, the products of nanotechnology research play an increasingly important role in the lab and clinic [476, 450, 457, 68].

It isn't so much a matter of *whether* nanotechnology will deliver implantable whole-brain scanning, but *when*, and, unfortunately, predicting when is complicated. On the one hand, it is possible to make some rough predictions about when we can expect a chip with a transistor count of 10,000, peak power consumption less than 5 nanowatts and size small enough to enter the brain through the capillaries might be feasible — anywhere from five to ten years depending on the degree to which Moore's law can be sustained[16].

On the other hand, issues of toxicity and heat dissipation pose problems likely to delay approval for use in humans, not to mention the consequences of implanting millions of computer chips in your brain. In this section, we look at what is possible now and what we believe will be possible in the medium- to longer-term horizon. As a thought experiment, we sketch a not-entirely-implausible technology for whole-brain scanning to focus attention on some of the technological challenges. Think of a brain-scale cellular network with recorders playing the role of mobile phones, local reporters that of cell towers, and relays that of phone-company switches or networked computers. [17] The purpose of this exercise is to explore the boundary between science fact and science fiction in order to better understand the promise and peril inherent in this rapidly evolving technology.

The technology we envision requires several types of implantable devices operating at different scales. At the smallest scale, we imagine individual cells instrumented with recorder molecules that sense voltage levels, proteins, expression levels, and perhaps changes in cell structure. There already exist micro-scale MEMS

[15] Specifically, we know of no practical method to induce exocytosis of large amounts of DNA that doesn't involve killing the cell.

[16] Here we are assuming that power consumption per device will continue to fall by half every 1.5 years, at least for the next five or so years. Koomey's law [227] describing this trend was based on performance over the last six decades but there are some worrisome issues projecting forward, especially when one considers the energy consumed in charging and discharging capacitance in transistors and interconnect wires and the static energy of devices in terms of controlling leakage current [424]. If progress is stalled or seriously retarded, the prospects for keeping to our ambitious time line will obviously be negatively impacted.

[17] It may one day be feasible to install a network of nanoscale wires and routers in the extracellular matrix. Such a network might even be piggybacked on the existing network of fibrous proteins that serve as structural support for cells, in the same way that the cable companies piggyback on existing phone and power distribution infrastructure. At one point, we discussed the idea of using kinesin-like "wire stringers" that would walk along structural supports in the extracellular matrix to wire the brain and install a fiber-optic network.

lab-on-a-chip devices employing scalable technologies that should serve to produce nanoscale variants as the fabrication techniques mature [475, 229, 266, 97]. These recording devices would transmit signals to be read by nanoscale reporting devices floating in the surrounding extracellular fluid or anchored at strategic locations in the extracellular matrix.

Imagine one type of reporting device distributed throughout the brain so there is one such device within a couple of microns of every neuron. Each such reporter is responsible for picking up the signals broadcast from recorders in its immediate vicinity. These reporting devices might directly transmit information to receivers located outside the brain but we anticipate this approach running into power and transmission limitations. Instead, suppose that these local reporters participate in a heterogeneous network of devices allowing for different modes of transmission at different scales.

The local reporters would forward information to a second type of reporting device more sparsely distributed and responsible for relaying the information received from the local reporters to external receivers not limited by size or power requirements. These relay devices could number in the thousands instead of millions, be somewhat larger than the more numerous and densely distributed reporters, and be located inside of the dura but on the surface of the brain, within fissures or anchored on the membranes lining the capillaries supplying blood to the brain and the ventricles containing cerebrospinal fluid.

Recorders might use optical signalling or even diffusion to transmit information to local reporting devices. The local reporters might employ limited-range photonic or near-field technologies to forward information to the nearest relay device. Finally, the relays multiplex information from many reporters and forward the result using microscale RF transmitters or micron-sized optical fibers connected to a "Matrix"-style physical coupling.

Relays located on the surface of the brain would have more options for power distribution and heat dissipation, and, being larger than local reporters, they would stand a better chance of having enough room to integrate efficient antennas. The RF transmitters would have no problem with scattering and signal penetration, and, because there are orders-of-magnitude fewer relays than local reporters, it is more likely that the available frequency spectrum will be able to supply sufficient bandwidth. Two-way transmission would enable sensing and the ability to excite or inhibit individual neurons.

Ignoring the potential health risks associated with such an invasive technology, most of the components mentioned above only exist in science fiction stories. That said, something like the above scenario is well within the realm of possibility in the next twenty years. Moreover, one can easily imagine less-invasive versions for medical applications in which the benefits outweigh the risks to patients. Indeed, we are already seeing clinical trials for related technological advances that promise to restore mobility to patients with severe disabilities [23, 140, 422].

Let's consider each of the components in our *Gedankenexperiment*, focusing on the reporting problems. The recorders associated with individual cells might encode data in packets shunted into the extracellular fluid for transmission to local reporters by diffusion. Suppose the packets are the size (˜50 nm) and diffusivity (˜10^{-8} cm^2/s) of vesicles used for cellular transport and suppose there are a million local reporters distributed

uniformly throughout a human brain. It would take about a day on average to transport a packet from a recorder to a nearby reporter [239].

A biologically-motivated alternative to decrease transit time and reduce the number of reporters might involve equipping each vesicle with a molecular motor or *flagella* providing an effective diffusivity of around 10^{-5} cm^2/s [61, 453]. With only 1000 reporters evenly distributed throughout the brain, it would take on average less than 2 hours for these augmented vesicles to traverse the necessary distance. Packing enough information into each vesicle in the recorder and unpacking, decoding and relaying the information in the reporter fast enough to support the necessary bandwidth is quite another challenge, but leveraging the molecular machinery of DNA replication and transcription might offer a possible solution — see Appendix 2.3.13.

Each reporter would transmit this information to the closest relay device on the surface of the brain. Depending on the application, we might want to distribute more relay devices on the surface of the cortex and thalamus than elsewhere. The surface area of an adult brain is around 2,500 cm^2. Given a thousand relays, each relay is responsible for a couple of square centimeters. Assuming a million local reporters proportionally distributed, each relay is responsible for forwarding information from a thousand local reporters. We would need each reporter to transmit up to $2.5 \times \text{sqrt}(2)/2$ cm = 1.76 cm.

Unfortunately, we can't position the reporters farther than about 500 μm and still retain the benefits of optical frequencies: low absorption to minimize tissue damage, high frequency to optimize bandwidth. Pushing aside these challenges, suppose there exists a frequency band in which scattering and absorption are low in brain tissue and relay devices along with their associated antennae are small enough to allow implantation.

If we allow for 10 bits per millisecond per neuron on average assuming only a fraction of neurons will have anything to report at any given time, 50% overhead for packetization, addressing, error correction, etc, then we arrive at 15 Gb per second as a bound on the channel capacity required of each relay. In any given millisecond, about 1% or 10,000 of the million reporters will be communicating with their respective relay — assume that neurons fire at 10 Hz and so only 10 milliseconds of every 1000 actually have a firing peak and require a signal.

For high-enough carrier frequencies, suppose we can use a modulation scheme like orthogonal-frequency-division-multiplexing (OFDM) [244], splitting up a congested bandwidth into many narrow-band signals. Then, assuming a reasonably powerful and compact source of Terahertz radiation, we could again apply OFDM and give each relay its own 15 GHz of bandwidth without inhibiting transmission — see Appendix 2.3.12.

As an alternative to relying entirely on the electromagnetic (EM) spectrum, researchers have proposed an approach in which the sub-dural reporters communicate with and supply power to a class of (implanted) reporters using ultrasonic energy [380]. The implanted reporters are referred to as *neural dust* and we'll refer to an individual implanted reporter as a *dust mote*. In this proposal, dust motes are also responsible for recording measurements of the local field potential (LFP) that the dust motes then transmit to the sub-dural reporters. The dust motes are approximately cube shaped, measuring 50 μm on a side, and consist of a piezoelectric device and some electronics for powering the mote and transmitting LFP measurements.

A 100 μm mote designed to operate at a frequency of 10 MHz, $\lambda = 150$ μm has ˜1 dB attenuation at a 2 mm depth — the attenuation of ultrasound in neural tissue is ˜0.5 dB/(cm MHz). A comparable electromagnetic solution[18] operating at 10 GHz, $\lambda = 5$ mm has ˜20 dB attenuation at 2 mm. There are other complicating factors concerning the size of the inductors required to power EM devices and efficient antennae for signal transmission — displacement currents in tissue and scattering losses can increase attenuation to 40 dB in practical devices. The proposed approach posits that electrophysiological data might be reported back via backscattering by modulating the reflection of the incident carrier wave used to supply power.

The small size of the implanted devices also complicates directly sensing the local field potential. The electrodes employed by electrophysiologists measure local field potentials at one or more locations with respect to a second, common electrode which acts as a ground and is located at some distance from the probe. In the case of LFP measurements made by dust motes, the distance between the two electrodes is constrained by the size of the device such that, the smaller this distance, the lower the signal-to-noise ratio — see Gold *et al* [146] for an analysis of extra- and intra-cellular LFP recording strategies for constraining compartmental models. The S/N problem can be ameliorated somewhat by adding a "tail" to each dust mote thereby locating the second electrode at some distance from the first one situated on the main body.

It is hard not to be a bit cavalier in fending off the technical challenges faced in trying to read off anything approaching the full state of an active brain, but the main purpose of our thought experiment is to acquaint the reader with some of these challenges, convey a measure of just how audaciously hard the problem is, and give some idea of the wide range of techniques drawn from many disciplines that are being brought to bear in attempting to solve it [269].

Noting that any of the nanotechnologies discussed would require some method of "installation" and assuming that our encouraging the development of "designer babies" with pre-installed recording and reporting technologies would be awkward and likely misunderstood, we came up with the following simple protocol for installing a generic nanotechnology solution along the lines of our *Gedankenexperiment*:

1. modification of individual cells to express biomolecular recorders might be accomplished using a lentivirus — unique among retroviruses for being able to infect non-dividing cells — or alternative viral vector and recombinant-DNA-engineered gene payload; while not a nanotechnology solution *per se*, this sort of biological approach will likely remain the best option for instrumenting individual neurons for some time to come,

2. distribution of local reporting devices might be accomplished using the circulatory system and, in

[18] Note that the speed of sound — approximately 500 m/s in air and 1500 m/s in water — is considerably slower than that of an electromagnetic signal — exactly 299,792,458 m/s in a vacuum and close enough to that for our purposes in most other media. The relatively slow acoustic velocity of ultrasound results in a substantially reduced wavelength when compared to an electromagnetic signal at the same frequency. Compare for example a 10 MHz, $\lambda = 150$ μm ultrasound signal in water with a 10 Mhz, $\lambda = 30$m EM signal. An EM signal of this wavelength would be useless for neural imaging; the tissue would be essentially transparent to the signal and so penetration depth would be practically unlimited, but there would hardly be any reflected signal and the size of an antenna necessary to receive such a signal would be prohibitively large — on the order of half the wavelength for an efficient antenna. A comparable EM solution for neural dust [380] would therefore be closer to the 10 GHz frequency provided in the text.

particular, the capillaries[19] in the brain as a supply network with steps taken to alter the blood-brain-barrier using focused ultrasound [188] during the installation process; once in the brain the reporters could be anchored in the extracellular matrix using techniques drawn from *click chemistry* [223],

3. pairing of neurons with their associated reporter devices might be accomplished using IPTG (isopropyl β-D-1-thiogalactopyranoside), tamoxifen, or other methods of controlling gene expression to achieve pairing; the associated machinery could be bundled with the lentivirus payload or introduced using a separate helper virus; this sort of *in situ* interface coupling between biological and nanotech components is likely to become common in hybrid solutions, and, finally,

4. installing a grid of relay devices on the surface of the brain without a craniotomy is likely to remain challenging until such time as nanotechnology develops nanorobotic devices capable of navigating in tissue either autonomously or via some form of teleoperation; in the meantime, there may be applications in which such a grid could be installed, at least on the cortical surface employing the same procedure used for intraoperative electrocorticography [377].

It is worth noting that a more natural method of "installation" might indeed involve a developmental approach patterned after the self-assembly processes governing embryonic development, but so far our understanding of molecular self assembly is limited to the simplest of materials such as soap films and, even in these cases, nature continues to surpass our best efforts.

The development of technology for building nanoscale communication networks [61] could propel our fictional brain-computer-interface into the realm of the possible. We're already seeing prototypes for some of the most basic components required to build communication networks, including carbon nanotubes to construct RF [221] and optical antennas [212] and multiplexers fabricated from piezoelectric nanomechanical resonators that could enable thousands of nanoscale biosensors to share the same communication channel [365]. It is possible to build a fully-functional radio receiver from a single carbon nanotube [195] and the possibilities for optical communication employing more exotic technologies from the field of plasmonics [378] are even more intriguing. However, as promising as these demonstrations may seem, practical developments in this arena appear to be some years off.

One area where nanotechnology is likely to have a big impact is in computing hardware. Logic circuits and non-volatile memory fabricated from carbon nanotubes are among our best hopes for sustaining Moore's law beyond the next few years. The prospects are excellent for smaller, faster, lower-power devices based on carbon nanotubes with possible spin-off benefits for neuroscience. Carbon nanotubes can be conjugated with DNA to render them biocompatible and it may be possible to embed these composites in cell membranes to be used as voltage sensors and to excite and inhibit individual neurons, thereby providing a practical alternative to silicon and other semiconductor materials that are toxic to cells.

[19] Needless to say it will be tricky shrinking the reporter technology to a size small enough that it can pass easily through the capillaries without inducing a stroke.

2.3.7 Technology Investment

Commercial opportunities and industry involvement related to a scientific endeavor can help to drive the development of enabling technologies and accelerate progress. In the case of the Human Genome Project, commercial interest lagged scientific progress in the early days, but once the business opportunities became apparent industry, involvement accelerated with collateral benefits for science in the form of better, cheaper and faster sequencing and related tools and techniques [100, 88]. In this section, we briefly survey some of the opportunities that might drive investment and spur innovation in neuroscience.

Companies that offer products and services relating to scalable neuroscience abound. There are companies like FEI, Gatan, Inscopix, 3Scan and Zeiss that specialize in electron microscopy, stains and preparations, and automation for serial-sectioning. Those that primarily serve the research community like Neuralynx and UNC Joint Vector Laboratories by providing electrophysiology tools for research including hardware and reagents for optogenetics. Industrial products and services like those offered by Nanoimmunotech including off-the-shelf and special-order nanoparticles and technologies for joining (conjugating) nanostructures, dyes, biomolecules. There are even grass-roots open-source communities like Open Optogenetics and Open EEG that create communities and offer educational resources and open-source tools and protocols. And then there is the incredible array of companies large and small that provide reagents, cell lines, hardware, software and services to the medical community.

Medicine has long been a driver for technology with ample capital funding to underwrite the development costs of new technologies. Assays for analyzing diseased tissue, tumors, cell counts, DNA sequences all have their neurophysiological counterparts, and, in the case of the brain, there is a felt need for lower-cost, non-invasive, readily-accessible diagnostic tools. The incumbent technology providers will likely maintain their technical and market advantages, but big companies tend to be slow to innovate and generally fail to offer incentives to their engineers to take on high-risk, high-payoff projects. Bulky, expensive hardware like MEG, MRI, PET, high-end EEG and ultrasound imaging and FUS equipment offer significant competitive challenges to the small company. That said there are opportunities in the synthetic biology and nanotechnology arenas to transition research ideas to products that could, in time, challenge even these markets.

In the near-term, companies like Neurosky, Emotiv, InterAxon, Zeo and Hitachi are pursuing BCI opportunities that leverage inexpensive, portable, non-invasive, off-the-shelf technologies such single-chip multi-channel EEG, EMG, NIRS, eye- and gaze-tracking, microfluidic immunoassay chips, etc. to provide tools and consumer devices for meditation, entertainment, sleep management, and out-patient monitoring. Near-infrared spectroscopy (NIRS) is a good example of a relatively inexpensive, non-invasive technology for measuring functional hemodynamic responses to infer neural activation — correlated with fMRI BOLD signals — which could be integrated with wearable displays like Google Glass.

One target of particular consumer interest is the development of personal assistants that know us better than we do, can help us to calibrate our preferences and overcome instinctive biases to make better decisions, and allow us to monitor or even exercise some degree of control over our emotional states to overcome anxiety or depression. Building such intimate software assistants and accompanying sensor suites that are

comfortable and even fashionable is challenging. Moreover a successful product will have to cope with the fact that the class of sensors that are practical for such applications will offer only rough proxies for those emotional and physiological markers likely to be most useful in understanding our moods and preferences.

There is reason to believe that existing options for sensing can be combined using machine learning to predict the relevant psychophysical markers. The obvious approach would be to create a training set using more expensive technologies such as MRI, MEG, and higher-quality, multi-channel EEG. Signatures for healthy and pathological variants of many cognitive functions are now well established to the point where we can predict common brain states by analyzing fMRI images. Researchers have demonstrated inter-subject synchronization of fMRI responses during natural music listening, providing a baseline that could be used to predict musical preferences[2]. The development of activity based classifiers to identify functional areas in animal models [46] could could be extended to take ECoG or EEG data from humans and do the same.

There are opportunities for small startups savvy in user-interface and machine-learning technologies to partner with neurophysiologists and cognitive neuroscientists to build products for personalized medicine, assistive technology and entertainment. Researchers at New York University and Johns Hopkins have shown that concussion, dementia, schizophrenia, amyotrophic lateral sclerosis (ALS), autism and Fragile X syndrome (FXS) are among the numerous diseases with characteristic anomalies detectable using eye movement tracking technology [307, 369]. Researchers at Yale used fMRI [372] to enable subjects to control their anxiety through a process of trial and error resulting in changes that were still present several days after the training. The experimental protocol relied on displaying the activity of the orbitofrontal cortex (a brain region just above the eyes) to subjects while they lay in a brain scanner.

In the medium-term, there will be opportunities using transcranial magnetic stimulation, focused ultrasound, implantable electrical and optogenetic arrays for patients with treatment-resistant anxiety, depression, stroke, head-injury and tumor-related tissue damage, neourodengenerative diseases, etc. Here again partnering with researchers, sharing IP with academic institutions, and finding chief executives and venture capital partners experienced with health-related technologies will be key. Longer-term opportunities for better prosthetics, cognitive and physical augmentation, entertainment, etc. are the stuff of science fiction but could arrive sooner than expected if key technologies mature more quickly than anticipated.

As for the considerable promise of nanotechnology, venture capital firms interested in this area might want to hedge their bets by dividing investment between (1) non-biological applications in communications and computing where biocompatibility isn't an issue and current technologies are up against fundamental limitations, and (2) biological applications in which the ability to design nanosensors with precisely-controllable characteristics is important.

Regarding (1) think in terms of quantum-dot (QD) lasers, photonics, and more-exotic-entangled-photon technologies for on-chip communication — 2-D and 3-D chips equipped with energy-efficient, high-speed interprocessor communication supporting dense core layouts communicating using arbitrary, even programmable topologies. Regarding (2) there is plenty of room for QD alternatives to natural chromophores in immunofluorescence imaging [276], voltage-sensing recording [274], and new contrast agents for MRI [319]. Advances

in precisely controlling QD properties will help to fuel the search for better methods of achieving biocompatibility.

There's another category of technology development that is more speculative but worth mentioning here. A number of companies like Brain Corporation, Evolved Machines, Grok — formerly called Numenta, IBM and Vicarious have embarked on projects to build systems that make use of ideas from neuroscience to emulate brain-like computations. Their business model is based on developing computational architectures patterned after the brain to solve practical problems in anomaly detection, visual perception, olfactory recognition, motor control and autonomous vehicle navigation.

There is a long history of neuroscientists using existing simple electrical, mechanical and computational models to explain neural circuits [277, 443, 399, 48]. It's not clear that we know enough about the basic principles governing real brains to engineer artificial brains. Indeed it would seem that if those principles were widely known they would have been incorporated into state-of-the-art computer-vision and machine-learning systems given that those fields include leading experts in computational neuroscience, but that is manifestly not true [104]. There is another possible approach to unraveling the secrets of the brain that holds out more promise and that dovetails with the focus on this report on developing new scientific instruments to record neural activity at the scale of whole brains.

Even the near-term technologies that we have discussed in this report promise to provide unprecedented detail across significantly larger populations of neurons than possible previously. Moreover with new tools from optogenetics and better optical methods for delivering light to deep tissues we now have the capability of selectively activating and silencing individual neurons in a given circuit. Optogenetic tools are particularly convenient for probing neuronal sensitivity, mimicking synaptic connections, elucidating patterns of neural connectivity, and unraveling neural circuits in complex neural networks [69]. It will soon be possible to instrument and record from neural circuits in awake, behaving subjects exposed to natural stimuli. The potential for hypothesis-driven science to discover the underlying principles is enormous, and the prospects for data-driven exploration may be at least as promising.

Engineers routinely apply machine-learning algorithms to fit models with thousands or even millions of parameters to data. Typically the data defines the inputs and outputs of a function that we would like to infer from the data and then apply to solve a problem such as analyzing an image — the input is a numerical array of pixel values — and determining if it contains an instance of a particular class of objects — the output is true or false. However, it is also possible that the function is realized as a physical system and the inputs and outputs correspond to measurable parameters of that system. In this case, the objective is generally to infer some function that reproduces the behavior of the physical system or at least the observable inputs and outputs in the data.

An investigator interested in inferring such a function might start with a hypothesis couched in terms of a family of models of some explanatory value, with success measured in terms of the fitted model accounting for observed data [213, 65]. Alternatively, the investigator might be satisfied with a family of models expressive enough to capture the behavior of the target system but of little or no explanatory power and a set of

experiments demonstrating that the fitted model accounts for the data, including held-out data not included in the training data. This sort of relatively opaque model can be employed as a component in a more complicated model and the resulting compositional model might prove to be very interesting, both practically and scientifically.

For example, a number of scientists believe that the neocortex is realized as a homogeneous sheet of anatomically-separated, computationally-similar units called *cortical columns* [297, 298]. If so, then if we were to infer the function of one cortical column in, say, the visual cortex, and then wire up a sheet of units implementing this function, then perhaps the composite sheet would exhibit the sort of behavior we observe in the early visual system. This particular example is quite challenging, but the same principle could apply to a wide variety of neural circuits, e.g., the retina, believed to have this sort of compositional architecture [472, 366].

A team knowledgeable in large-scale machine learning might partner with one or more labs involved in recording from neural circuits and share data and expertise to essentially *mine* the data for algorithmic and architectural insight. Products amounting to black boxes realizing useful adaptive strategies and pattern recognition behaviours would provide substantial value in a wired world in desperate need of such behaviors exhibiting the robust character of biological systems. Scientists scour the world for novel genes and microorganisms that exhibit useful behavior and it is this trove of natural technologies that hold out such promise for the next generation of neural recording instruments [359]. It may be that these same instruments will help to reveal an equally valuable algorithmic windfall for the companies with the wherewithal to harvest it.

2.3.8 Acknowledgements

This report is the culmination of nine months of research into the topic of scalable neuroscience and, in particular, the prospects for developing new scientific instruments to accelerate research in the brain sciences. During the first six months, a small band of us including Ed Boyden, Greg Corrado, David Heckerman, Jon Shlens, Akram Sadek and, toward the end of this period, Yael Maguire and Adam Marblestone collaborated to explore the space of relevant technologies and identify promising opportunities. We also enlisted the help of a larger group of scientists many of whom participated in CS379C at Stanford and provided consulting to the students working on final projects. In the final three months prior to releasing this document, the students in CS379C listed as coauthors on the report joined the effort identifying additional opportunities and analyzing those considered most promising. We are particularly grateful to Ed Boyden, Kevin Briggman, David Cox, Mike Hawrylycz, Yael Maguire, Adam Marbelstone, Akram Sadek, Mark Schnitzer, Jon Shlens, Stephen Smith, Brian Wandell and Tony Zador for joining in our class discussions. We would also like to thank Arjun Bansal, Tony Bell, Kwabena Boahen, Ed Callaway, John Donoghue, Bruno Madore, Bill Newsome, Bruno Olshausen, Clay Reid, Sebastian Seung, Terry Sejnowski, Dongjin Seo, Fritz Sommer and Larry Swanson for timely insights that served to adjust our perpectives and shape this report.

Section 2.3.8

2.3.9 Leveraging Sequencing for Recording

Overview

The BRAIN initiative, Brain Activity Mapping (BAM), Human Brain Project, and others like it seek to gather the activity and structure of many neurons; this data would help test comprehensive models of the brain and design better treatments for nervous system diseases, e.g. Huntington's, Dementia, Pain, and others. The human brain has on the order of 10^{11} neurons, 10^{14} synapses, over a hundred neurotransmitters, and dozens of genetically defined cell types among other parameters [17], which traditional technologies — microscopy, electrophysiology, electrode arrays, and others — are ill-equipped to deal with. This magnitude and specificity of data is normally overwhelming with without serious compromises: experiments are invasive and subject viability is reduced. In contrast, DNA sequencing is used to measure millions to billions of molecules, is readily scalable (both reading, storing, and analyzing the data [256]), and always getting cheaper [143]. Due to this, next-generation sequencing technology is proposed to help connectome construction and activity recording. Doing this will require integration of several technologies, many of which are already available and ready to use while others still need some tweaking. Based on analysis of recent publications and currently available technologies, we are optimistic that sequencing can be used to produce connectomes and firing patterns in cell culture or simple model organisms in the near term (1-2 years) with use in rodent models in the medium term (3-5 years). However, a concern remains about the viability of this approach to study human connectomics without serious technical improvements, on both the science and policy side.

There are several proposals on how to use DNA sequencing to study connectomics and activity, we will focus on two and several technologies that would help in their implementation. To study the connectome, the plan is to individually barcode all neurons in a brain, allow the barcodes to spread to synaptically connected partners, ligate host and foreign barcodes, and sequence [471]. To record activity, one proposal is to encode the activity as errors in a DNA template [228]. These are fundamentally molecular biology challenges that need to be overcome, however the possibility exists that using nanotechnology [8] will help improve the reliability and experimental breadth of these technologies by taking advantage of the increasing returns in computing power and size reduction as epitomized in Moore's [294] and Bell's law [30] laws along with Intel's Tick-Tock model of chip architecture miniaturizationd [391].

Technical

Some of the technologies outline below are available (Recombinases, Super-resolution microscopy, FISH, etc.) while others still need to be fully developed (ion-sensitive and cytoplasmic DNA polymerases, induced DNA secretion, etc.). However, a timeline for when they can be integrated into a sequencing-based approach to BAM is of interest given the many parts needed to get the entire systems working.

Ion-sensitive DNAPs DNA polymerases (DNAPs) can be sensitive to ion concentration and a recent publication characterized a particular one, Dpo4, for its transfer function (relation between ion concentration and error rate) with Mn^{++}, Mg^{++}, and Ca^{++}. Dpo4 was only useful for discerning Mn^{++} and Mg^{++},

which are physiologically less preferred than Ca^{++} [473]. However, it is likely that ion-sensitive DNAPs can be found or created by mining microbiology literature and metagenomic searches[20] or through DNAP engineering (e.g. via directed evolution [135, 458, 122, 136]). For example, modifying DNAPs can yield increases in desired parameters, such as the addition of a T3 DNA polymerase thioredoxin binding domain to Taq polymerase that caused a 20-50 fold increase in processivity [99]. Further, a basic search of Polbase, a DNA polymerase database, shows that T7, Pow, and Pab Pol I all have very low error rates (lower than Dpo4); in addition, T7 has high processivity and is quite rapid—13,404 bases/min or 223.4 Hz. Given that networks can fire between 40-200 Hz, this should allow enough resolution to pick up spiking at rest or low frequency activity [264]. The availability of many DNAPs and the pressure to improve them within the molecular biology community — for use in PCR and other assays — indicates that an ion-sensitive DNAP could be created or found within 1-2 years and implemented in mammals in 3-5 years.

Cytoplasmic DNAPs Transcription in the cytoplasm would be preferred for activity-based measurements using DNA polymerases to avoid the different characteristics of calcium transients in the nucleus [42]. A special form of transcription that can occur in the cytoplasm has been observed and might be adapted for the ticker tape system.[85] Given the paucity of experimental data showing localization of DNAPs to the cell membrane — either via anchoring to transmembrane proteins, addition of GPI anchors, or other methods — it is unlikely that this component will be implemented until 3-5 years out unless a breakthrough in adapting cytoplasmic viral replication machinery, such as that of the Mimivirus, can be demonstrated.

Modified DNAP trangenetic mouse It takes at minimum around two years to make a trangenetic mouse line [95]. Given that ion-sensitive nor non-viral cytoplasmic DNAPs haven't been fully characterized and a cell line containing a barcode cassette has just been made[470] indicates that we are 3-5 years from obtaining a mouse model that natively expresses a modified DNAP, either constitutively or under control of specific inducible constructs (e.g. Cre, FLP, etc.).

Recombinases Cell lines with a *Rci*-based shuffling cassette stably integrated to allow random barcoding of neurons has already been created [470] and a randomized transfection library already demonstrated to work [317]. Full characterization in cell culture and invertebrate model organism is likely 1-2 years out while use with rodents is 3-5 years out (see transgenetic mice). This technology is unlikely to be used with humans in the near future.

2^{nd}/3^{rd} generation sequencers PacBio RS, 454 GS/FLX and Life Technologies Starlight[21] have the longest read lengths needed for activity-based sequencing. However, Starlight isn't available yet, PacBio has an extraordinarily high error rate of around 12%, and 454 GS/FLX series have position-dependent error rates, which are not preferred. While it is possible to implement the proposed BRAIN sequencing methods now, improvements in read length, error rate, position-dependent errors, and error type are needed to reduce the problem of biases. Because of the competitive environment and rapid pace of improvement in this area, we expect the needed technology is in the 1-2 year pipeline and if needed can be optimized specifically for these

[20] BLASTing for DNAPs that appear to have calcium binding domains or similar structural elements to DNAPs with known calcium sensitivity.

[21] Not yet on the market.

applications given adequate interest and funding.

Super-resolution microscopy STORM, PALM, STED, and other microscopy techniques have helped illuminate the fine structures of the cell and commercial systems are already available.[373] Integrating blinking fluorophores with GRASP, FISSEQ, and FISH for identification of synaptically connected neurons and *in situ* sequencing is possible in the next 1-2 years and has been previously supported to occur given adequate interest.[184]

FISSEQ/FISH/ISH-HCR These technologies have been improved over the past decade and incorporation into this system should be trivial [290, 29], with an estimate of 1-2 years. They each require slightly different approaches, but we envision FISSEQ or ISH-HCR winning out due to versetility in design of barcodes and ability to sequence error-strewn activity templates. Clear$^{\mathrm{T}}$ and CLARITY might provide a method of imaging without needing to arduously slice brains, but they are new techniques and at the moment not optimized. It will likely take 1-2 years given concerted effort and 3-5 years given parallel development to realize FISH or FISSEQ in 3D volumes without the need for slice work.

GRASP Another older technology that might be adapted for use in BRAIN sequencing [132]. The idea would be to synaptically couple two barcodes then sequence using FISH/FISSEQ and distinguish closely localized synapses using super-resolution microscopy. Skepticism remains about the ability to distinguish multiple points at the diffraction limit, but our calculations indicate that at an average of 1 synapse per cubic micron, this shouldn't be a large problem.

Exosome-based DNA secretion Cells secrete exosomes (small vesicles containing protein, DNA, and other small molecules) and they could be made to carry DNA containing connectome or activity data into the blood stream and out the renal system [41]. However, few mechanisms for experimentally inducing their secretion are known (calcium being one signal, but that would cause problems in a calcium detecting system). Technology to non-invasively alter nanoparticles exist for thermal, radio, and magnetic signals [185, 396, 32], so it may be possible to couple these technologies to induce DNA secretion to allow the subject to be kept alive during readout. Due to the multiple systems that need to be put into place, this is optimistically a 3-5 year outlook.

Engineered proteins These would be proteins that respond to ultrasonic, thermal, radio, magnetic, and other non-invasive signals. This has already been demonstrated for ion channels, but awaits conjugation to DNAPs or other proteins for integration into BRAIN sequencing systems [185, 396, 32]. Given that pieces of the technology exists we estimate 3-5 years for proof-of-concept and 5-10 years for expression in rodents.

Multiplexed probes The possibility exists that multiplexing delivery of sequencing chemicals and readout of the signal is possible, seeing that it has already been done with neural recording and drug delivery [379]. For example, the drug delivery channel would uptake small samples of surrounding fluid and use a nanodevice with a DNAP molecular imprinted onto its surface. Binding and sequencing of local nucleotides could be offloaded to a nearby sequencer or a clever method could take advantage of changes in surface conductance upon binding[22] of specific segments of the DNA strand, measure the change, and read this out as a measure

[22]For example, how binding in surface plasmon resonance (SPR) changes reflectivity.

of the current nucleotide being sequenced (given some previously defined standard) [268]. Given that we can now either offload the heavy lifting to already available DNA sequencers and because the probes have already been fabricated, preliminary results could be seen in brain slices within 1-2 years.

Nano-sequencers Inserting sequencers *in vivo* that don't need the addition of expensive reagents, high-cost library construction, and other add-ons associated with parallel sequencing[143] would yield benefits both from near real-time data acquisition and reduced complexity. Detecting proteins [67] and DNA [229, 266, 97] using micro-technologies already exists and nanopore technologies should allow these to be scaled down [209, 475, 175]. Already, theoretical calculations for DNA nano-sequencers has been demonstrated and the possibility of integrating this with other nano-scale read-out technologies (e.g. RFIDs or OPIDs) could allow sequencing without the DNA needing to leave the system [286]. Further, this would also improve the false-negative rate as DNA can be degraded or altered during transport out of the body. The fact that theoretical calculations for detecting base-pair differences sans pyrosequencing have been done but devices don't yet exist peg this as a 5-10 year technology.

Predictions

The following table summarizes the best estimates of when key technologies discussed in this proposal will be implemented for use in DNA sequencing for connectomics or activity recording. These times represent optimistic estimates for proof-of-concept in cell-culture then onto animal models (add an extra 2-3 years). We do not consider use in humans as that is 10+ years away for sequencing-based technologies, partially because gene therapy also needs to advance to a point where the constructs designed for use in model organisms can be easily transfered with drastically diminished risk for neuronal death induced by over-expressing viral vectors and other complications. Because of this long timeline, human implementation is not directly relevant to BRAIN goals in the short/medium term for this set of technologies (in contrast to fMRI, EEG, and other medically approved, non-invasive technologies).

Technology	Area	Years	Notes
Ion-sensitive DNAPs	activity	1 to 2	[a]
Cytoplasmic DNAPs	activity	3 to 5	
Modified DNAP transgenetic mouse	activity	3 to 5	
Recombinases	connectome	1 to 2	
$2^{nd}/3^{rd}$ generation sequencers	both	1 to 2	[b]
STORM/PALM/STED	both	1 to 2	
FISSEQ+2D slice	both	1 to 2	
FISSEQ+CLARITY/Clear^{T}	both	3 to 5	[c]
FISH+2D slice	both	1 to 2	[d]
FISH+CLARITY/Clear^{T}	both	1 to 2	
GRASP	connectome	1 to 2	
Exosome-based DNA secretion	both	3 to 5	
Multiplexed probes	both	3 to 5	
Engineered proteins	both	5 to 10	[e]
Nano-sequencers	both	5 to 10	[f]

[a] DNA polymerase sensitive to Mn and Mg exist and are mentioned in Zamft *et al* [473].
[b] Need to see improvements in error rates and read length.
[c] The ability for 3D imaging to obtain diffraction limited sampling in unknown at present.
[d] This would be more useful for connectomics where a known set of barcodes is possible. One could envision using the percent binding of a template probe to determine activity.
[e] Ultrasonic microbubbles are currently being considered for gene delivery but few proteins are known to respond to it specifically.
[f] Based on estimates from current lab-on-a-chip devices.

2.3.10 Scalable Analytics and Data Mining

Overview

We present some estimates of the scale at which we would be able to record from neurons simultaneously. The treatment here is data modality agnostic, i.e. we may have calcium imaging data, data from optical sensors, or data from direct electrode measurements of neuronal voltages. The focus here is more on the potential of computer analytics to identify functional patterns with the measured data. The goal may be to identify functional collections of neurons or achieve super resolution imaging based on the fact that the activities of nearby neurons are correlated.

A central problem in deducing function/structure from measurements is the issue of identifying which cell gives rise to a certain measured signal. Any measured waveform is contaminated by waveforms from nearby cells. Spike sorting algorithms try to rectify this by separating individual waveforms (from different neurons) from their linear combination. One can envision that these spike sorting algorithms combined with sophisticated side-information will enable us to record from more neurons than is possible now with electrodes. We provide an example below to illustrate the kind of scaling possible and the development time

scales involved.

Of course such techniques would invariably be guided by the functional models of the brain and the quality of measured data. In the following sections we review some of the technologies and algorithms which show great promise in helping us scale our readout problem. While most of the discussed technologies are currently possible only in animal models because of the need for novel signal measurement devices (requiring genetic enginering or highly invasive electrode recordings), one can imagine that the signal processing algorithms used here would be data agnostic and hopefully will carry over even to non invasively gathered human brain data.

Technical

Most of the methods in this section focus on signal recording of neural activity, either in the form of direct measurements of voltage levels, or indirect measurement of the calcium concentration levels using genetically encoded calcium indicators (GECIs). Although the latter usually suffer from lack of resolution (mainly temporal) some GECIs like GCaMP3 and/or GCaMP5 [192] offer much faster kinetics (i.e. greater temporal resolution) and greater stability across interesting timescales (i.e. greater readout duration). As mentioned later, such properties are crucial for increasing the readout duration from neurons, and may serve as convenient replacements for micro electrode arrays for high throughput brain signal measurement. These signals are used in conjunction with sophisticated microscopy (laser or fluorescence) and/or sophisticated algorithms/models to perform parameter estimation. The hope is that prior knowledge of the structures or processes generating the measured signals would either help in the refinement of the estimated parameters or in scaling up the number of signals read. In fact if the sizes of the neurons are large enough (e.g. some hippocampal and cerebellar cells) and if the functional regions are stable enough (across behavioural states and time scales as is the case with hippocampal place cells), such techniques offer practical ways to scale up the number of neurons one can image.

One of the main issues in making sense of measured signals from the brain is identifying which cells they come from. While rate of spiking was long considered to be critical for understanding neural function [398] , it was slowly realized that other waveform properties (like relative spike timing patterns) may also be important in encoding neuron function. This is where spike sorting algorithms come in. These try to separate linear superposition of signals into waveforms for each neuron. Algorithms like PCA (principal component analysis), ICA (independent component analysis), and particle filter (sequential Monte Carlo methods) have been successfully used in some settings to recover spiking patterns of individual cells. While these algorithms are not free from artifacts, these techniques in combination with other modalities like cell-body segmentation do offer powerful ways of inferring connectivity patterns and/or functional classification from signals [193].

Predictions

In general the exact reconstruction techniques used depends on the area of the brain we are imaging and the signals or parts of the brain we are interested in. Techniques like the ICA (independent component analysis)

have been used to good effect in understanding biomedical signals (EEG, MEG, fMRI) [435]. Of course, such analysis tend to be based on some assumptions about the underlying signal generation process, in particular about measured waveforms being a linear combination of several independently generated waveforms. While this in isolation may not correspond to realistic models of neural function, it does offer useful insights especially when combined with different techniques [203]. The number of sample points needed however scales as the square of the number of independent components we are trying to recover. One can think of applying a similar concept in the analysis of neural signals. The independent components would correspond to different functional units. Combining measurements from different approaches may help to annotate reconstructed signal with e.g. location information from microscopy.

Using ICA, we generally need of the order of n^2 data points to be sampled in order to estimate the mixing coefficients of the n independently firing neurons. Thus for approximately $80-100$ neurons in a cortical minicolumn or $7-50$ neurons in a cerebellar microzone (an anatomical and functional collection of neurons), we need to be able to record on the order of 10^4 data points. Since neural connectivity patterns or functional evolution can be significantly affected by timeshifts of a few milliseconds [272], we need at least one millisecond temporal resolution. This entails that we record at least for $10s$ or more. While this is not hard with direct electrode recordings (which often have good temporal resolution), the use of indirect (e.g. calcium imaging) techniques often need sophisticated signal processing to achieve the same (e.g. to deduce precise spike times from calcium levels). Thus not only do we need to record from many neurons simultaneously, we also need to be able to record for a longer duration per neuron. Due to plasticity of synaptic connections and functional associations and the photobleaching effects (or instability) of the indicators, this may be a challenge in some settings. However, in some cases mentioned below where there is stability of the functional units across time or behavioural states, the problem is solvable using the approaches mentioned here.

We note in passing that while these ideas are representative of the gains that can be achieved, they are not the only ones considered in the literature. The problem of spike timing inference has been looked at from other perspectives also. For example, [441] uses particle filtering to deduce spike times. With simulated data they conclude that using models can outperform standard estimation methods like Wiener filters. However, for brevity and to convey representative estimates of how much we can scale, we henceforth focus on specific work.

We consider the work in [299], and look into the evolution of similar techniques. By using techniques like independent component analysis together with cell body segmentation, the authors in [299] could recover signals from $>$ **100** Purkinje cells. By analysing the correlation patterns, they could also identify microzones with sharp delineations (width of one Purkinje cell). This level of precision was not achieved by direct measurement methods at the time it was published due to the coarse (approximately 250μm) inter-electrode spacing. Future work in [106] used a combination of two photon imaging and its correlation with virtual reality patterns to achieve functional imaging of hippocampal place cells at cellular resolution during virtual navigation. By using a different microscopy technique, researchers in [137] were able to obtain simultaneous recordings from $>$ **200** Purkinje cells simultaneously across different microzones. This was scaled to $>$

1000 hippocampal pyramidal cells in [480]. Of course, the absolute numbers are dependent on which part of the brain we image and the specific techniques we use, but the numbers do reveal the power of more refined models of neural activity and improved signal recording capabilities in understanding and better characterising the functional state of the brain.

In the remaining part of the discussion we present some estimates for the potential of better algorithms in helping us understand other functional units of the brain which may not necessarily be as stable as the microzonal structures in the cerebellum. A topic of intense research in this is the cortical minicolumn [62]. Proposed in [296] in 1957 this has led to significant research in identifying groups of neurons as a functional unit, instead of focussing on single neurons. It talks about functional organization of cortical neurons in vertical columns, the idea being that they would be activated by the stimulation of the same class of peripheral receptors. This idea is the basis for much of future work. For example, the Blue Brain Project [271] aims to simulate the human cortical column on the Blue Gene supercomputer. The basic workhorse is still the Hodgkin Huxley model, simulated on a supercomputer to reveal interesting macrostructures. While there is strong evidence for cortical microcolumns being important units of neural function and pathways, they do show some sort of plasticity, i.e. the neurons from different cortical columns may organize themselves into units which change with time and functional stimulus [62]. Of course having the model is not enough. One needs to be able to fit the free parameters in the model to the empirically measured data. This is one of the main focus areas for the Blue Brain project also. Readouts from patch clamp techniques or other MEA (microelectrode array) recordings are used to model the different neurons. Other techniques like spectral clustering (which is similar to PCA) offer some insights [114].

In humans, these minicolumns range in length from $28\mu m$ to $40\mu m$. While electrodes offer ways to record from neurons in different parts of the brain simultaneously, they usually suffer from poor spatial resolution (but good temporal resolution). However, some of the latest developments in electrode technologies offer a solution to that. Using the resolution offered by the latest 3D optogenetic arrays, we can get close to 150 μm [482], but other techniques like fluorescence microscopy [137] offer feasible ways of imaging cortical columns in a level of detail which would be enough to resolve neurons to single cell precision.

In short these techniques in conjunction with microscopy and other data acquisition techniques offer scalable and non destructive ways of increasing spatiotemporal resolution (as compared to using them in isolation e.g. microelectrodes which have good temporal resolution but poor spatial resolution or traditional imaging tools which generally have poor temporal resolution-although some recent work [154] promises sub millisecond accuracy). In any case, these techniques can help not only to reconstruct from measured data, but also to guide data acquisition. Incorporation of available side-information about the neurons we are interested in can offer powerful ways of scaling up the number of neurons we are able to image.

2.3.11 Macroscale Imaging Technologies

Overview

Nuclear magnetic resonance (NMR) imaging modalities are currently the most promising technologies for recording macroscale measurements, with resolution on the order of millimeters and seconds, used in the analysis of the major functional areas and the white-matter pathways connecting those areas in awake, behaving human subjects. Magnetic resonance imaging (MRI) sensitive to quantitative NMR tissue properties, diffusion, and blood oxygenation are currently the tools of choice for studies of normal and pathological behavior in the field of cognitive neuroscience as well as clinical diagnosis.

Diffusion tensor imaging (DTI) and quantitative MRI (qMRI) have already been used in *in vivo* human studies to quantify fascicle and tissue development, and therefore cognition and behavior. DTI models the distribution of diffusion directions of water protons as a tensor, providing measures of the apparent diffusion coefficient (ADC) and fractional anisotropy (FA). Identical lobes of a gradient in the diffusion-sensitive direction are applied separated by a 180 degree pulse and a temporal wait so that diffusing spins, unlike stationary spins, are not fully re-phased and thus contribute attenuated signal. Quantitative mapping of the proton density (PD), which is proportional to the amount of water, and the T_1, which measures the spin-lattice relaxation constant and thus quantifies the interactions between protons and their molecular environment, can be used to measure Macromolecular Tissue Volume (MTV), the non-water volume in the voxel, and the Surface Interaction Rate (SIR), the efficiency of a material's energy exchange with water protons [281]. Since water preferentially diffuse along axons, DTI enables tractography of the fascicles and since approximately 50% of the macromolecules in the white matter are myelin sheaths, MTV measurement provides an indication of axonal diameter and myelination. Although there is no dynamic activity mapping, the changes in structural properties give an indication of function due to long-time-scale development and plasticity of the brain. Specifically, the rate of change of FA, MTV, and SIR in the posterior corpus callosum, the arcuate fasciculus, and the inferior longitudinal fasciculus have been shown to correlate with reading ability [444]. In addition, changes may be used to diagnose and evaluate diseases; for example de-myelinating disorders such as multiple sclerosis can be identified using MTV measures.

Functional MRI (fMRI) enables the coarse spatial and temporal localization of neural activity through the proxy of the hemodynamic blood oxygen-level dependence (BOLD) response. After neural activity, ions must be pumped across cell membranes for repolarization. This necessitates blood flow to the region, resulting in elevated oxygenation levels in the 2 to 3 millimeters surrounding active neurons 2 to 6 seconds after they fire. Oxy-hemoglobin is diamagnetic like most substances in the body, while deoxy-hemoglobin is paramagnetic. Therefore, the increase in oxy-hemoglobin levels results in a decrease in the magnetic susceptibility differences between the blood and the surrounding tissue. Since the variation in the proton resonant frequencies correspondingly decreases, T_2* lengthens, and the image appears more intense in the active regions. Taking advantage of this indirect coupling of a magnetic spin parameter to synaptic activity, it is possible to map the brain activity in response to motor and cognitive tasks, thereby enabling the better understanding of these functions at the macroscopic level. Although the seconds-scale response time is much

slower than many neural processing dynamics, the information from fMRI has still proven very valuable. For example, fMRI has facilitated the identification of the visual field maps in the human visual cortex and improved the understanding of the perception and function of the visual system [446].

The advantages of MRI stem from its noninvasiveness, its endogenous contrast, and its scalability, and therefore the fact that it can be used in human studies. Although MRI does not have neuron-level resolution, it does have the ability to measure more macroscopic variables that reflect properties of the local population of neurons and how these local populations interact. It thus has the potential to address the missing length scales in neuroscientific knowledge — while neurons and synapses are reasonably understood, we do not have a good understanding of the microarchitectural units in the brain and especially how these units communicate, interact, and work together in a global network. Therefore, since we do not yet have the ability to understand the complexity of individual neuron data across the entire brain, courser spatial-scale information is valuable for increasing the understanding of the brain and is scalable in terms of both acquisition time and data processing and storage. MRI in particular can be used to study the integration of signals across brain circuits and the development of the white matter, and therefore cognition, over time and experience. Especially in these areas of research, recording from humans is necessary, as smaller, less-developed animals lack the extensive pathways and complex cognitive functions of humans [445]. Therefore, the promise of MRI primarily stems from its ability to be used on humans to study the global brain network.

Technical

Currently, scan times of 8 to 12 minutes were necessary to achieve 1.5 mm^3 to 2 mm^3 resolution at 1.5 T and ˜1 mm^3 resolution at 3.0 T for qMRI and DTI [281]. The long duration of MR image acquisition posed a challenge in prior studies, preventing the study of cognitive development before the age of eight [445]. Although fast sequences have been used for both qMRI and DTI, the acceleration factors from parallel imaging, if any acceleration was used at all, were relatively conservative [281]. Therefore, there is still a lot of room for improvement in the use of parallel MRI (pMRI) systems to facilitate the greater understanding of the development of the brain at the macroscale. A system with four receiver coil elements and a corresponding four will be utilized in the very near future to scan six to eight year olds [445].

The fMRI signal, stemming from susceptibility effects of blood oxygenation, is typically very weak. Therefore, signal averaging is necessary to achieve satisfactory image quality, and high main field strengths are utilized to increase the signal strength, and therefore the signal-to-noise ratio. Acquiring the signal multiple times to facilitate averaging, however, increases the scan time by a factor of the number of averaged signals. The acceleration from parallel imaging, therefore, can keep the image acquisition duration reasonable, especially for pediatric imaging. Especially at the high field strengths used for fMRI, the acceleration factors obtained from pMRI are high. In addition, parallel imaging ameliorates the tradeoffs incurred from increasing the main field strength.T_2* generally shortens as field strength increases, causing the signal to dephase quickly during readout [144]. Since the BOLD response is rapid, however, fast sequences with extended readout times such as echo planar imaging (EPI) are used for fMRI. Therefore, the shortened readouts

provided by parallel coils are particularly valuable in fMRI to avoid the decay of the signal. Parallel imaging has already been successfully used for functional imaging [101]. Gains from parallel imaging have not been exhausted, however, and improvements in pMRI over the next 2-3 years are predicted to significantly contribute to functional neuroimaging.

Coil arrays with up to 32 elements which can achieve up to an eight-fold improvement in scan time without significant degradation of the image quality. Using current RF coil hardware, if the acceleration factor is increased further, the sensitivities of the coil elements used for each measurement overlap spatially and the measurements are no longer necessarily independent in the presence of noise. It thus becomes difficult to solve the inverse problem of reconstructing the image from the set of acquired signals and coil sensitivities as the effective rank of the relevant matrix (resulting from the signals and sensitivities) decreases. In addition, the number of elements in RF coil arrays are limited by a signal-to-noise ratio (SNR) penalty. Since smaller coils are more sensitive to the surface than to deeper structures, reconstructing the image at the center of the brain requires summing the small signals from all of the array elements. In doing this, noise is also received from all of the coils (rather than primarily from the nearest surface coil, as is the case for superficial structures), and the SNR degrades to become comparable to that of a body coil [308]. As better RF coils are designed, improving the spatial selectivity and uniformity (over the limited volume) of the elements, coil arrays with 256 elements and an acceleration factor of 32 can be expected within the next 2-3 years. This improvement may enable the testing of subjects as young as 2 or 3, which would, for example, allow the study of cognitive development that begins earlier than reading, as well as the development of motor skills, which occurs at too young of an age to be studied with current technologies. In addition, in other applications where fast scanning is not essential, the scan time improvement can be traded in for SNR and resolution improvements.

Improvements in image acquisition time and signal to noise ratio, as well as the application of the most advanced current technology to neuroscience studies, therefore, will very probably contribute to the understanding of the brain's structural development and function within the next three years.

Ultrasound

The lack of an acoustic window into the brain is a significant roadblock to the use of ultrasound or noninvasive neural imaging in humans. A back-of-the-envelope illustration of this uses the characteristic impedance mismatches between air, bone, and tissue and therefore the reflection coefficients, the frequency necessary for desired resolution, and the attenuation coefficient at this frequency (using the ˜1 dB/MHz/cm loss rule of thumb). These determine the intensity necessary for the signal to be above the electronic noise floor at the penetration depth into the brain desired. The intensity at the focus, combined with the absorption coefficient of the tissue, gives the specific absorption rate (SAR), and the bioheat equation can be used to determine the increase in temperature from the SAR. Using the Arrhenius damage integral, the temperature function can be converted into an equivalent thermal dose at 43 degrees C, which gives an indication of if the tissue has been damaged via coagulative necrosis. In addition, the intensity of the pressure waves can be compared to thermal

and mechanical limit indices to evaluate the safety. Due to the acoustic properties of the skull, ultrasound may be more suitable for neural stimulation using low-intensity, low-frequency ultrasound, or therapeutic ablation using high intensity focused ultrasound (e.g. of the thalamus for essential tremor treatment) rather than imaging. The main contribution of imaging to this field, therefore, may be MR, which provides thermometry measurements and good contrast, to guide and monitor the use of focused ultrasound. An alternative may be to surgically insert transducers, but, in doing so, some of the main advantages of imaging, noninvasiveness and therefore the ability to conduct human studies, are lost. Since the main challenges to using ultrasound in neuroscience is not temporal or spatial resolution, but depth penetration/SNR and heating issues, despite recent, near-term, and intermediate-term improvements for rapidly scanning tissue at high temporal (dynamic) and spatial resolution, it is likely not most promising technology for brain imaging.

2.3.12 Nanoscale Recording and Wireless Readout

Overview

We discuss in the following section the use of micron-scale, implantable optical devices for wireless readout. We present an optical identification tag (OPID), an analog of the conventional Radio-frequency identification tag (RFID), as a sensor for neuronal monitoring of the firing of axon potentials as well as a communication means with other devices. After a brief description of the structure of OPIDs, including their necessary size and their components, we present two technological schemes for wireless readout via nanotechnology, each leveraging the use of OPIDs and potentially realizable in the next 4-8 years. In each, we envision a region of the brain (potentially the entire brain) where an OPID is placed next to each neuron such that it can record, in real time, the firing of the axon potential.

The first readout system would utilize larger, yet still micron-scale, implanted RFID chip reporters to relay information from the sensor OPIDs to devices external to the brain. Each reporter would communicate locally with the sensors nearest to it, and would then transmit an aggregated signal, containing the data of its local sensors, to an external receiver, where the data could be processed. This system would allow for real-time monitoring of the firing of large regions of neurons. The OPIDs could be installed via the methods described in the introduction, and the larger RFID reporters would be surgically implanted. The second readout system is similar to the first, but uses an insertable fiber-probe to communicate with the sensors. Each fiber probe would contain hundreds, or even thousands of chips capable of communicating with the sensors, and would remain connected to some external circuitry. However, instead of relaying the information wirelessly, each of these reporter chips would send data along the probe itself via electronic circuits.

Technical

With Moore's Law holding over the past many decades, transistors have reached dimension sizes of 22 nm and are projected to decrease to 10 nm by the year 2016. As such, the design of an RFID with micron dimensions has been achieved to well within the dimensions of neuronal cells [58], given their size range of

4 to 100 μm. In fact, the latest in RFIDs can fit inside the larger ones. Assuming a linear scaling of RFID size with transistor size, we can expect a 10,000 transistor RFID to decrease to about 10 μm $\times$ 10 μm $\times$ 5 μm by the year 2016.

One type of RFID, proposed by Yael Maguire at Harvard, is an optical frequency RFID, or OPID, measuring 10 μm $\times$ 10 μm $\times$ 5 μm and containing ˜10,000 transistors that would sit either inside a neuron or immediately adjacent to it. Each OPID will contain a neuronal voltage sensor [328], sensitive enough to detect voltage levels on the order of millivolts, with a time-resolution within 1 millisecond. The OPIDs will operate while under the constant illumination of the reporters and utilize small optical components capable of communication with either larger RFID chips or fiber-probes. When the OPIDs sense axonal firing, a CMOS circuitry layer with a CPU will store this information, and at specified intervals dynamically change the load impedance of their optical components to modulate backscattered signals. Maguire suggests the powering of such a device using a high-efficiency solar cell. Wireless power transfer via magnetic induction and glucose powering are two other avenues for powering these devices [236]. State of the art glucose energy extraction has achieved 1.0-3.3 μW/cm^2 or 1.0-3.3 $\times$ 10^{-2} pW/μ m^2 [160]. Transistors today operate at picowatt scales, so either the glucose extraction efficiency or the transistor operating power need to increase/decrease by about 3 orders of magnitude. Extrapolating Moore's law, we can expect glucose powering to become feasible in 16 years.

RFIDs as Reporters: We now consider the first system, in which OPID sensors communicate with local RFID reporters that send information outside the brain. These OPIDs will modulate the backscatter of light sent by the RFIDs. We choose optical communication in the wavelength range of 600 nm to 1400 nm, the *biological window*, due to its low absorption coefficient [353] of 0.1-0.5 mm^{-1} and its high frequency, enabling low-loss communication with substantial bandwidth. This choice results in a tradeoff with scattering. Light scattering at these wavelengths is quite strong, and leads to a very small coherent penetration depth. Grey matter in the brain suffers from a scattering coefficient of approximately 10 mm^{-1} [465], corresponding to a signal-to-noise ratio (SNR) of e^{-10} for two objects separated by 500 μm. White matter has an even greater scattering coefficient of approximately 30 mm^{-1}, corresponding to a SNR of e^{-30} at the same separation distance. Given this constraint, an OPID in grey (white) matter will need to communicate with a separate reporter within a distance of about 500 μm (150 μm).

As RFIDs decrease in size, RFID-based implantable biomedical devices continue to decrease as well. Pivonka *et al* [330] report a 2 mm $\times$ 2 mm wirelessly powered implant with both a communication channel in the low GHz frequency range and a magnetic induction powering range of 5 cm. The device itself spans only 600 μm $\times$ 1 mm $\times$ 65 nm, with the larger dimensions being the result of the coil used for wireless power transfer. A mere 500 μW is all that is needed to continuously power the device, and this takes into account the power that goes into the locomotive motion of Pivonaka's design - a stationary device would require substantially less power. This powering distance range, along with this power threshold, allows for access to the majority of the brain. In fact, the powering distance is the main limiting factor in how deep into the brain an RFID can be placed. Communication distances are not an issue provided the chosen frequency band has

low extinction in brain tissue, and provided that the RFID is large enough to contain a dipole or folded dipole antenna functional in this frequency band [125]. Therefore, it is reasonable to expect, in the next 4-8 years, a decrease in the size of the magnetic induction coil relative to the wavelength to the point where the same range of 5cm could power a coil of 500 μm $\times$ 500 μm, given the decrease that has occurred since wireless power transfer was derived [236, 330].

Having addressed the issues of powering RFIDs and OPIDs, as well as defining frequency windows of communication for each, we next turn to the interaction between the sensor OPIDs and their reporter RFID counterparts: We assume the implantation of a 10 μm $\times$ 10 μm $\times$ 5 μm OPID for every neuron in some region of interest (ROI) of the human brain, yielding a sensor density of 8×10^4 sensors / mm^3. This leads to a 4% increase in the volume of the ROI, or identically a 4% decrease in extracellular fluid. For illustrative purposes we consider a region of grey matter, where a sensor-reporter pair can be distanced up to 500 μm apart. Assuming the insertion of RFIDs on the order of 500 μm $\times$ 500 μm $\times$ 30 μm spaced every 1 mm, then each reporter would be linked to all sensors within a 1 mm^3 rectangular volume surrounding it. Given that the firing rate of a neuron is approximately 10 Hz, then we can expect 800 spikes/ms on average.

A given OPID will need to transmit its own identity in the form of some serial number, timing information about the firing of the axon potentials, and an additional amount of overhead accounting for error correcting codes, redundancy, etc. If the OPID transmits a packet of data every time its neuron fires, it will need to transmit about 100 bits per packet, assuming 12 bits to contain firing information, 37 for an OPID identification number, and just over 50 for overhead. For a 1 mm^3 ROI, this corresponds to 8×10^5 total signals sent per second, (sps) and 8×10^7 bits/second (bps). While the bps is small, the sps could lead to congestion at the reporter end, similar to bad cell phone reception in a densely crowded area. Suppose that, instead of an OPID transmitting every time its neuron fires, it stores the data in a buffer and transmits it every second, yielding a sps of 8×10^4 sps. The necessary buffer could be achieved with Maguire's 10,000 transistors, in 4 years in accordance with Moore's law. Reducing the sps by another factor of ten to 8×10^3 sps would require 10 times the buffer size. This is accomplishable, for the same-sized device, in 8 years time.

Next we must consider the issue of bandwidth between the RFIDs and OPIDs. As mentioned above, the biological window provides an 800 nm wavelength (λ) range in which transmission can occur. As an illustrative example, we consider frequency-division multiplexing between the OPIDs. That is, each OPID is assigned, by the RFID, a specific frequency band in which it will communicate, in order to distinguish between signals. In communication systems, the Q-factor $Q = f/\Delta f$ is the ratio of the frequency of transmission to the half-max bandwidth, and is a measure of frequency selectivity. For 800 OPIDs around 1 RFID each utilizing 1nm λ of bandwidth, each OPID will need to transmit with a photonic device offering $Q = 1.05 \times 10^3$. For 4×10^4 OPIDs per RFID in a volume of 1 mm^3, photonic devices with $Q = 52.5 \times 10^3$ will be needed. While this may be achievable with quantum dots in years to come, this is possible to achieve at the moment with very thin photonic crystal slabs. Specialized photonic crystal cavities have demonstrated ranges from $Q = 45{,}000$ [6] to $Q \geq 1{,}000{,}000$ in the wavelengths of interest [15]. Another solution, also

proposed by Maguire, is the use of an LCD screen in the OPID that can modulate the backscatter of an impinging electromagnetic (EM) wave. It has been shown that by applying a small voltage, the reflectance and efficiency of an LCD display can be boosted to $\geq 95\%$ [225].

Electronic Fiber Probes as Reporters: Using RFIDs as reporters introduces the additional challenge of wireless relay from the RFIDs to external devices. RFIDs will work well for small ROIs, but the bit rate transmitted from them scales linearly with volume. In the limiting case of the entire brain and each OPID transmitting at the above 8×10^7 bps, the total data rate will be 2×10^{14} bps, or 2 Terabits/sec. Instead of attempting to handle an enormous amount of data wirelessly, it may be more feasible to replace the RFIDs with thin electronic fiber probes, on the order of 500 μm in diameter. Suppose a ROI in the brain had OPIDs attached to every neuron, and further suppose that a series of probes, in parallel, were surgically implanted into this ROI. By spacing the probes with the same characteristic spacing of the RFIDs (~1 mm^3), and placing RFID-like chips inside the probes, at this same characteristic spacing, it would be possible to read data from the neurons and transmit information electronically up the probe instead of dealing with it wirelessly. The drawback to this system is the added volume needed in which to place the entire probe.

2.3.13 Hybrid Biological and Nanotechnology Solutions

DNA Sequencing Implants

Nanotechnology is poised to offer much benefit in sequencing DNA, and this benefit can be harnessed for brain activity mapping. There are many proposals for using DNA to record information from neurons, such as recording synaptic spikes and mapping the connectome — see Appendix 2.3.9. We expect other information (such as topography) also will eventually be recordable in DNA. Once information is recorded in DNA, however, getting the information out still presents a formidable challenge. One possibility is to encapsulate the DNA in a vesicle, which would migrate through the extracellular fluid to DNA sequencing chips inside the brain, at which point the DNA would be sequenced and the information would be read out digitally.

The migration of the vesicles from neurons to the chips creates a surprising amount of difficulty. These vesicles would probably have diffusivities similar to other such vesicles, or around the order of 10^{-8} cm^2/s, causing them to be rather slow [239, 178]. If 1 million sequencing chips were evenly distributed throughout the brain, it would take around a day for these vesicles to travel the distance needed to reach the chips. A better solution (one that would both decrease the time and number of chips) would be to equip these vesicles with molecular motors. Fitting these vesicles with molecular motors such as flagella could give them effective diffusivities of around 10^{-5} cm^2/s [61]. With only 1000 chips evenly distributed throughout the brain, it would take on average less than 2 hours for these augmented vesicles to traverse the necessary distance. 1000 chips is a small enough amount that they could be implanted manually, though we expect such a procedure to be automated. The vesicles additionally would contain functional groups to allow them to target to these chips. A large number of each of these vesicles would need to be released for each corresponding strand of DNA to ensure that at least one makes it to a chip. Once the vesicle makes it to the chip, it would release the

DNA for sequencing.

Sequencing the DNA could be performed quickly by nanopore technology. Nanopores consist of small holes in a membrane (such as in silicon or graphene). As DNA is threaded through this nanopore, voltage readings on the membrane indicate translocation events that correspond to specific base pairs, allowing for fast sequencing. Currently, each pore can read 300 bp/s (base pair per second) [224]. A chip containing an array of 100×100 nanopores reads 3 Mbp/s. Assuming DNA sequencing follows its exponential increase in speed (doubling time of about 0.8 years), this leaves us with speeds of around 200 Mbp/s and 12 Gbp/s for these chips in 5 and 10 years [404]. Assuming 1000 chips in the brain and estimating that a bit of data might take tens of bp to encode, then in 5 years this system would be able to hand 10 Gbit/s, and in 10 years it could handle 1 Tbit/s. The 5 year option presents 1 bit for every 10 neurons per second for readout, which is likely too small for mapping activity for anything other than short experiments, but could be used to determine other information, such as cell types or connectomics. By 10 years from now, this technology would have 10 bits per neuron per second. In theory, this would allow for mapping the activity of 10% of the neurons in the brain with 10 ms resolution, or 1% of the neurons in the brain with 1 ms resolution. If trends continue, all 100B neurons could be readout with 1 ms resolution in about 15 years. Finally, this information would be read out from the chips using fiber optics. With each chip processing 10 Mbit/s in 5 years and 1 Gbit/s in 10 years, this is clearly within the range that fiber optic wires could handle.

Förster Resonance Energy Transfer

Förster Resonance Energy Transfer (FRET) is a phenomenon that can be exploited to optically map the activity of the brain. FRET can occur when two chromophores are close to each other; the donor chromophore transfers its energy to the acceptor chromophore by dipole-dipole coupling, leading the acceptor to fluoresce. If the chromophores are not close, the donor fluoresces at a different frequency. Optical imaging techniques can therefore be used to determine if the chromophores are within a certain distance. Placing these chromophores either on different molecules or specific locations of the same molecule allow one to determine if the two molecules are interacting or if the one molecule has undergone a conformational change, and this can be harnessed for brain activity mapping. The two most promising approaches for using FRET are a genetic engineering based approach and a nanotechnology based approach.

Genetic engineering has already led to some successes in using FRET for mapping brain activity. For instance, Cameleon is a genetically encoded calcium indicator (GECI) that in the presence of calcium ions undergoes a conformational change which increases the FRET effect [291]. Cameleon can therefore be used to visualize synaptic spikes. GCaMP5s (GECIs similar to Cameleon) are possibly the most advanced of such genetically encoded indicators and have been used to image the firing of over 80% of the neurons in an entire zebrafish brain *in vivo* at 0.8 Hz and single-cell resolution [5]. With each iteration of the GCaMP molecule, there have been a few large improvements of properties (such as three-fold increases in contrast

from GCaMP3 in 2009 to GCaMP5G in 2012) and a number of smaller improvements [192] — see Appendix 2.3.14. Extrapolation of these trends indicates that these molecules will undergo incremental improvements in coming years. The genetic engineering approach to FRET has the advantages of being more mature than the nanotechnology approach and being capable of implementation in a relatively easy, nontoxic manner. However, the requirement for genetic engineering does present a potential barrier for use in humans.

The nanotechnology approach to FRET typically involves quantum dots (QDs), nanoparticles that confine excitons in all three spatial dimensions and thus exhibit only a few allowed energy states. QDs can be used in FRET as either the acceptor or the donor (or both), and have been used to image action potentials [302]. QDs have many beneficial properties for use in FRET. For instance, QDs have broad absorption spectra with narrow emission spectra, and varying the size of QDs substantially varies these spectra. Therefore, they are bright, tunable, and have high signal to noise [476]. The biggest detriment against using QDs for FRET is probably their toxicity. There are currently efforts to reduce toxicity by coating QD in polymers or functionalizing them with ligands, but these efforts so far have only seen partial success [134]. Another hurdle will be introducing the QDs into the brain. Focused ultrasound can be used to temporarily disrupt the blood-brain barrier, allowing for the QDs to cross into the brain [449].

The genetic engineering based approaches to FRET seem likely to dominate for at least the next five years, due to their relative maturity and problems QDs face. Nanotechnology ultimately has more to offer, and probably will overtake the bioengineering approaches after toxicity and delivery hurdles are solved, likely between 5 and 10 years from now. For both of these approaches, imaging in mammals and in particular people presents issues largely due to the opaqueness of the brain. Microendoscopy or related technology could be used to overcome these problems — see Appendix 2.3.15.

Carbon Nanotube Neural Stimulation

In addition to mapping brain activity, nanotechnology has the potential to control the activity of the brain with great precision. The realization of such capabilities has many market and social incentives. In addition to treatment or cures for neurological disorders, fine control promises benefits in many endeavors, including gaming, learning, and augmentation. Furthermore, controlling the activity of the brain will be instrumental in mapping the activity of the brain, as control will allow us to discern causality instead of simply correlation. Carbon nanotubes (CNTs) are one class of molecule that can be exploited for this control.

One such scheme involves using DNA to interface CNTs with ion channels. CNTs can be separated by length using centrifugation [123]. These different length CNTs can then be wrapped in specific DNA sequences that target particular chiralities of CNTs, such that each sequence corresponds to CNTs of a particular length and chirality [423]. These resultant complexes (DNA-CNT) are nontoxic and able to cross the blood brain barrier [108]. These DNA-CNT would be fitted with strands of DNA on their ends such that the strands target particular ion channels. This targeting will probably initially require the ion channels to be genetically engineered so that the DNA-CNT have an easier time homing in on them, but we envision that eventually the DNA-CNT could be functionalized to specifically bind to non-engineered ion channels. By

differing the targeting DNA strands depending on the length of the CNT, different length DNA-CNT will target to different types of ion channels [364]. CNTs oscillate when irradiated with light (with oscillation depending on length of the CNT), and this oscillation in turn activates the ion channel [364]. By using different frequencies of microwave light, we can target different length CNTs and thus different ion channels [364]. Using only semiconducting CNTs and not metallic CNTs — which are dependent on the chirality — prevents the CNTs from heating up under this irradiation [364].

This process would have a lot of the same capabilities as optogenetics, but would also have certain advantages. For one, this method would allow for many more types of ion channels to be independently activated. This is because the narrow range of light usable for optogenetics would allow for only a handful of ion channels to be independently activated without much crosstalk. The carbon nanotubes, on the other hand, could be fabricated to different lengths such that a very large number of channels could be independently activated [364]. Additionally, microwaves can penetrate the skull, while visible light cannot, so this method can be used without implants. Furthermore, we might be able to target DNA-CNT to ion channels without genetic engineering, thus circumventing a major obstacle for use in people. Of additional note, DNA-CNT is slowly broken down by the body, so this technique would not be permanent.

2.3.14 Advances in Contrast Agents and Tissue Preparation

Overview

The technologies presented in this section all fall into the category of near-term opportunities with potential for incremental progress over the longer term. Non-invasive means for detecting neural structure and activity is paramount for both basic neuroscience and clinical therapies. Novel contrast agents for MRI and photoacoustic tomography (PAT) present technologies that can be developed to reliably gain large scale (whole-brain) structure and activity information. A holy grail for all imaging techniques is to enable identification of specific molecules. Several avenues for using contrast agents to identify molecules are being explored in these techniques. All of these technologies are available for use in animal models. Potential for clinical use is noted where applicable.

Genetic engineering has matured to the level of complexity development. DNA can be deterministically created and inserted into a genome. There already exist genetically encoded indicators for a variety of neurotransmitters. Finally, optogenetics provides a tool to optically excite or silence neurons. All of these pieces have brought genetic engineering to a point of complexity development. The leading question here is, how can one combine these techniques to learn more about the biology of the brain? As these techniques have already been developed, incremental progress will proceed with a tilt toward complexity of implementation. Finally an important class of genetically encoded indicators are calcium indicators, also known as genetically encoded calcium indicators (GECIs). Currently GECI's present important information regarding neuronal activity, but do not present activity on the level of a single action potential. The current state of the art GECI, Gcamp5, exhibits great improvement over previous versions and incremental advancement is expected.

Novel tissue preparation techniques like CLARITY [87] and Clear^{T} [237] present new ways for researchers to optically probe prepared tissue. Previously, researchers needed to cut small portions of the brain, stain, and image this small area. With CLARITY and Clear^{T}, one can stain large portions–an entire mouse brain — entire brain and image separate areas confocally. These techniques also allow for washing and restaining of the same tissue. Incremental progress will proceed in refining the technique, and research groups will ramp up use of the tool.

Technical

Focused Ultrasound (FUS) has been combined with photoacoustic tomography for deep (3mm in brain tissue) and accurate imaging of blood brain barrier (BBB) disruption [450]. In this work, gold nanorods (AuNR) were used as a contrast agent for PAT to image the time dynamics of BBB disruption by FUS. Gold nanorods are particularly well suited as a contrast agent for photoacoustic tomography (PAI) due to their tunable optical properties and the potential for gene delivery [79]. Additionally, gold nanocages have been explored as contrast agents with capability for drug delivery [462].

Photoacoustic Tomography produces images by using ultrasound and the transduction of absorbed photons into heat and subsequently pressure. It has been used in mice for functional imaging of hemodynamics without the use of a contrast agent. It has the advantage over MRI of imaging both oxygenated and deoxygenated hemoglobin, two values correlated with cancer and brain function [451]. Additionally it can be combined with contrast agents for molecular targeting. It is compatible currently available molecular dyes [251] that have potential for approval for clinical use. Finally, PAT has potential for clinical use as it is non-invasive and can image intrinsic absorbers. To date, PAT has achieved imaging on many scales — from organelles to organs — with the capability of imaging up to 7 cm into brain tissue with sub-millimeter resolution [449]. PAT can be implemented in a variety of different manners; it has been used to image blood flow (as in doppler photoacoustic tomography) [448], gene expression [250]. Many implementations of PAT are intrinsic and noninvasive. Thus, they present opportunities for human studies and clinical use. PAT has already been commercialized for preclinical use through VisualSonics and Endra. It is predicted that PAT usage will increase within 1-2 years for preclinical applications and rapid adoption and commercialization for clinical use will follow upon FDA approval.

Magnetic Resonance Imaging provides another method for non-invasive measurement of neural activity. Bioengineering of contrast agents serves a powerful method to develop contrast agents; one could imagine building a versatile indicator for MRI that is similar to the green fluorescent protein (the workhorse of molecule targeted optical microscopy). Magnetic nanoparticles such as super paramagnetic iron oxide when conjugated with calmodulin can be use as a calcium indicator for neural activity. Directed evolution has been used to develop a dopamine indicator based on a heme protein [385]. Finally, enzyme based contrast agents combining MRI contrast with in situ chemical processing. An early example visualized β-galactosidase expression via enzymatic hydrolysis of a Gadolinium substrate [261]. Given the mature level of this technology incremental development of molecular targeted contrast agents will occur.

GCaMP, a popular GECI, has become an indispensable tool for neuroscience research. The state of the art, GCaMP5, cannot reliably detect sparse activity–one test setup provided detection of single action potentials (APs) 26% of the time [192]. Previous versions of GCaMP had negligible detection of single APs. Dynamic range has also rapidly increased GCaMP2, GCaMP3 and GCaMP5H exhibited $\Delta F/F = 5.1 \pm 0.1$, 12.3 ± 0.4, and 158 ± 12 respectively. Finally, calcium affinity has also increased over previous versions of GCaMP. The rapid progress of the GCaMP family of indicators and the recent release of GCaMP6 (on AddGene) indicates that this technology will continue to develop the potential for calcium imaging of sparse activity. Furthermore, GCaMP has been used in mice, drosophilia, zebrafish, and C. elegans and can be presented as a viral vector or can be expressed transgenically. GCaMP is already commonly used in neuroscience laboratories to image neural activity. However, many advancements remain. Thus we place this in the short term advancement and implementation category with the hope that reliable detection of single action potentials arrives within 3-5 years.

The recent introduction of CLARITY and Clear$^{\mathrm{T}}$ to the neuroscience community presents a great tool and will quickly revolutionize experimental neuroscience [87]. The technique allows for multiple rounds of in-situ hybridization and immunohistochemistry. Combining CLARITY with optogenetics and/or calcium imaging can provide detailed information regarding activity, structure, and expression. Clear$^{\mathrm{T}}$ presents a powerful tool for developmental biology and interrogating. At the moment it is unclear in what ways the technique will be used. However, the latest tissue clarification techniques open new possibilities for research and will become commonplace in laboratories within 1-2 yrs.

2.3.15 Microendoscopy and Optically Coupled Implants

Overview

In this section we talk about implantable devices for activity recording in dense neural populations. The focus here will be particularly on recordings from deep brain regions as opposed to superficial layers that are easier accessible by other means. We first give the estimates for the scalability of microendoscopy approach. We then investigate avenues for development of next generation of implantable devices that can further boost the number of recorded locations and individual neuron cells per location.

Microendoscopy: Optical ways to probe and stimulate neural activity are gradually taking over traditional methods such as multi-electrode arrays. This is facilitated by great advances in GECI and more recently voltage sensitive fluorescent proteins. A number of fluorescence microscopy techniques utilizing these advantages have been employed to record neural activity. Among them: single- and two-photon excitation scanning microscopy [103], light-sheet microscopy [5] and others. Due to severe light scattering, imaging depth is limited to $\sim$500-750 um, so only superficial brain tissue is reachable for conventional microscopy techniques [455]. The inside of the brain machinery is, of course, as essential as the periphery and the main challenge is to obtain high-quality recordings from deep brain regions as well.

It has been shown that implantation of a carefully engineered microendoscope allows extending capabilities of standard microscopy techniques to regions arbitrarily deep in the animal brain. Microendoscope composed of several GRIN (Gradient Index) lenses relays the high quality optical image above the animal scull, which can be further accessed by conventional optics. Strikingly, activity from more than a thousand cells in a mouse Hippocampus can be recorded simultaneously using this technique [480]. Both single and two-photon imaging modalities are possible using GRIN endoscopes [216]. Thus, scaling the microendoscopy approach will be extremely useful to allow recording from multiple dense populations of neurons in several locations along the neural pathways in the brain. This should greatly facilitate studies of connectivity between different brain areas as other techniques do not provide this kind of resolution and coverage. Since most technological problems have been extensively debugged in single-endoscope experiments, it is reasonable to expect steps in direction of multiplication of microendoscopy recordings in the same animal. We believe that first attempts will be made in 1-2 year perspective and further increase in the number of simultaneous endoscope implantations will follow in 2-5 years. However, one of the main disadvantages of microendoscopy is its inherent invasiveness. This also becomes a limiting factor for recordings from too many endoscopes, since brain function can be substantially altered by implant. In the following technical section, we estimate the total number of simultaneously implanted endoscopes for deep-structure imaging not to exceed 20, which corresponds to recordings from $\sim$20000 neurons simultaneously. Though this number is just an order of magnitude larger than what is currently available with other technologies, it still opens unprecedented experimental opportunities, thanks to dense recordings at distant locations.

Fiber-coupled microdevice implants: Since scaling of microendoscopy recordings has its limitations, mainly due to high invasiveness, different approaches and technologies will emerge to keep up with the demand to obtain neural activity recordings at cellular level from more and more brain locations and more cells per location as well. In a general way we here discuss the advantages of using an optical fiber as an implant that can solve two major technological problems. First, sufficient power has to be supplied to the 'recorder' of neural activity. In case of ion or voltage sensitive florescent proteins excitation light has to illuminate neurons; in case of electrode arrays one needs to power all the electronics. In both cases optical power can be sent directly into the fiber. Second, optical fiber may have an extremely large bandwidth and serve to transmit optical signal encoding neural activity out from the brain.

Fiber implantation has been used in a variety of optogenetic experiments [31]. However, stimulation with a single fiber lacks spatial resolution. Some attempts to construct fiber bundles have been made [169] which is though as bulky but inferior in quality to microendoscopy. Recording fluorescent signal with a fiber itself is also not feasible as it does not carry spatial information, so optogenetics field is moving into direction of microfabricated multiwaveguide arrays [31, 482]. We suppose that fiber implantation may become extremely valuable if coupled to a miniature device located on its tip. We envision the following functions such microdevice needs to possess:

1. microdevice needs to be well interfaced with the fiber to form a single minimally invasive implantable unit,

2. monitoring by optical or electrical means activity from several hundred to several thousand of surrounding neurons,

3. interface with the fiber, namely encode the spatial information and neuron firing times/patterns into an optical signal that is transmitted back into fiber, and

4. efficiently manage optical power sent into the fiber from outside: direct incoming optical power to specific spatial locations (in case of fluorescence excitation) or convert it into an electrical signal to power built-in electronics.

In the following subsection, we propose several implementation options and performance estimates. In the 5-10 year range, this approach may allow recording a comparable to microendoscopy number of cells per locations while being much less invasive.

Technical

Microendoscopy: Typical microendoscope for imaging deep brain structure includes a micro-objective and a relay lens is $\sim$5mm long and $\sim$0.5 mm in diameter. This results in damaging of minimum 0.2% of total brain mass of a mice (typically 0.4-0.5g). This accounts only for the volume that has to be replaced by an implant. In reality, damage might be significantly larger due to immune response to a foreign body. A number of studies examined different parameters of electrode implants that affect the damage to the brain. Such factors as implant material and size as well as the insertion speed determine the number of damaged neuron cells [331, 418, 39]. Many of the findings can be extrapolated on microendoscope implants, however, to the best of our knowledge, no comprehensive studies have been conducted to quantify the effect of multi-endoscope implants from both microscopic and behavioral points of view. Though in some experimental protocols damage to certain areas of brain tissue is tolerable, one should not aim at more than 20 simultaneous deep-imaging sites, which will result in brain lesion of $\sim$4% by volume.

By extrapolating density of recordings from [480] we come to a conclusion that engineering advances can bring the microendoscopy to simultaneous recording from $\sim$20000 neurons deep in mouse brain. However, further increase of the number of endoscopes will potentially result in unreasonable complication of experiment and different deep-imaging approaches are needed. A system capable of handling and manipulation of such big number of microendoscopes in conjunction with simultaneous imaging would require significant engineering effort especially in miniaturization of optomechanics, though technology itself is available. We believe that due to extremely high interest and investment in the field, these issues may be solved in the nearest perspective, and functioning multi-endoscope imaging systems appear available to the neuroscience community in 1-2 years, and more advanced systems in 2-5 year perspective.

Fiber-coupled microdevice implants: The idea of a microdevice recording neural activity and coupled to an optical fiber is very generic and can be possibly implemented in a variety of ways and employing different technologies. For example, the sensing part of the device may consist of multi-electrode arrays or interface with different nanoscale recorders scattered in the brain (see above for details). Another option is to

record activity information in an optical way directly, e.g., building a microdevice with a sensor that preserves spatial information: a combination of micro-optics and CMOS sensors, for example. Ideally, the lateral size of the microdevice should be matched to the diameter of the fiber, while constraints on its length are not as strict. This should be done to ensure that the whole implant causes minimal damage to the brain.

According to [418], implants 50 μm in diameter lead to larger survival fractions of neurons compared to bigger implants. To be more specific, we assume that the fiber with cladding and a device have a diameter of 100 μm. This would result in 25-100 times less tissue damage than in case of typical microendoscope of the same length. A 50 μm or 62.5 μm diameter multimode fiber has a bandwidth of 10Gb/s, which is well above the bandwidth needed even for a very high resolution data one can be recording in a single location, e.g., 1000×1000 pixels image, 8-bit depth and 100Hz update frequency results in 0.8Gb/s.

An emerging field of Photonic Crystals (PhC) may greatly facilitate all-optical signal routing and processing on the microdevice itself. Efficient PhC-based waveguides, splitters and couplers and all-optical switches [197] have lateral footprint of $(\lambda/n)^2$, so a thousand distinct waveguides (one per recorded neuron) can easily fit laterally in the microdevice each addressing a distinct 'recorder' or a portion of the sensor. In terms of power consumption the fiber can transmit up to several watts of optical power, so the limitation would be set by the amount of power that brain can dissipate (typically several tens of mW per imaging location). PhC all-optical switches also have an advantage over electronic counterparts as they can work in sub-femtojoule per bit regime [310] (to power 100000 devices at 10 MHz one needs only 1 mW of power). The whole implanted device can be rapidly switched between input/output modes: fraction of the time for power input and the rest for recorded data streaming. Of course, the exact architecture of a microdevice needs to be carefully engineered, but achieving similar to microendoscopy numbers of recorded cells should be feasible within 5-10 years, while significantly reducing the invasiveness.

2.3.16 Opportunities for Automating Laboratory Procedures

Overview

We are interested in the applications of machine learning and robotics to automate tasks previously carried out by scientists and volunteers. This brings speed and consistency to experiments, but this also comes with questions and demands in error rates and efficiency. This is especially important in the brain readout problem, where an error rate of less than 0.01 percent can result in drastically skewed data once brought to a larger scale, eg. from a neuron to the connectome. In evaluating emerging technologies, we looked for improvements in scalability. The technologies chosen to focus on all minimize the risk of scalable error. Furthermore, recognizing scalability as both vertical and horizontally applicable, we weighed technical value on more than strides in the error rate. Considering horizontal scalability to be the technology's potential to expand into a human model and the primary obstacle to this being the invasiveness of the technology itself, we chose to focus on three technologies ranging in progression toward a noninvasive technique: patch-clamp electrophysiology, automated scanning electron microscopy, and high-throughput animal-behavior experiments.

Scanning Electron Microscopy (SEM): Scanning electron microscopy generates three-dimensional images using a combination of two-dimensional images generated by focusing an electron beam across the surface of a biological tissue sample and collecting data on the backscattered electrons. This technology has seen tremendous advances in the last few years and is at the forefront of today's imaging methods. Furthermore, as electron microscopy staining has shown to be successfully unbiased in the staining of membranes and synapses in a neuron, in principle, the technology has the potential to be quite successful in mapping neural networks. There are three technologies within SEM that are automated, and as their automation accuracy and resolution improve, they are all viable techniques in reconstructing neuronal connectivity. These are serial block-face SEM (SBEM), automated serial-section tape-collection scanning electron microscopy (ATUM-SEM), and focused ion beam SEM (FIB-SEM). These methods range in imaging resolutions and dimensions, but all have the ability to be automated in a way that is scalable to reconstruct dense neural circuits. Once the challenge of increasing spatial scope and resolution is addressed, the main challenges that remain for SEM in automation itself are increasing the imaging speed. It seems likely that in the next one to two years, acquisition will be fast enough to image an entire human brain in a reasonable amount of time (under one year).

Patch-Clamp Electrophysiology and Probe Insertion: Patch-clamp electrophysiology is a robotic tool to analyze the molecular and electric properties of single cells in the living mammalian brain. Automation of the patch clamp technique began in the late 1990s and current patch-clamp algorithms allow for the high throughput detection, electrical recording, and molecular harvesting of neurons. Recent advances, including a robotic process that allows for patch-clamp electrophysiology in-vivo, coupled with the ability to obtain information about the position or type of cellular structures being recorded indicate that this is a viable candidate for mapping neuronal connectivity. The major challenges to patch-clamp electrophysiology are throughput and the success rate of whole-cell recording. Nonetheless, there are significant breakthroughs currently being made in the field and these drawbacks will likely be resolved in the next five years.

High-Throughput Animal-Behavior Experiments: High-throughput animal-behavior experiments offer a means of studying human systems through an animal model. A challenge in using an animal model is acquiring and analyzing enough data to demonstrate that an animal model is an adequate comparison; therefore high throughput data collection using robotics is a viable way to expedite progress in this direction. Though animal-behavior experiments are less often considered at the forefront of viable technologies in mapping neuronal connectivity, development of new animal models exploiting characteristics of novel organisms may bring great advances in their parallels to the circuitry of certain parts of the human brain. The challenge of high-throughput animal-behavior experimentstion lies in being able to draw direct parallels from the animal to human models in brain circuitry, especially beyond the proven models in the visual system. In the next one to two years, expect to see an increase in high-throughput animal-behavior experiments as they are proving to be a viable option.

Technical

Scanning Electron Microscopy (SEM): *Increasing Imaging and Acquisition Speed:* Highly parallel SEM is a development whose implementation is highly feasible in the next one to two years. It is possible to parallel imaging across multiple microscopes by assigning each to separate imaging sections. This would increase imaging speed by over two orders of magnitude. Another feature commonly overlooked is the overhead amount of time it takes to section, load, and unload specimens. This can result in a per-section overhead of up to six minutes, which in the case of ATUM-SEM is over ten percent of the per-section imaging time. Automating parts of this overhead component is achievable within the next one to two years and would increase imaging speed up to ten percent. In addition, a recent study in SBEM used a novel technique for coating specimens in scanning to eliminate distortion in electric fields due to the accumulation of negative charge; the technology is applicable to other automated SEM technologies [419]. The overall result of such is the possibility to saturate camera sensors in a single frame, thus increasing system throughput. Coupled with the technological advances in image frame readout inherent in Moores law, camera acquisition speed may nearly double in the next one to two years [52].

Patch-Clamp Electrophysiology and Probe Insertion: *Increasing Throughput and Whole-Cell Patch Recording Success Rate:* Current automated patch-clamp electrophysiology technologies can detect cells with 90% accuracy and establish a connection with such detected cells about 40% of the time, which takes 3-7 min in vivo. It is also worth noting that the manual comparison was a success rate of about 29% [222]. Nevertheless, the low successful connection rate limits the amount of data being collected and in order to achieve more comprehensive recordings, the area may require a great deal more sampling. In addition, the 3-7 minutes of robot operation is over a small localized area: scaling this method up to the entire brain may prove to require a lot more time, especially in the overhead of setting the robot up. This would be an undesirable amount of time for an in vivo study to take place. There are advances in increasing the speed of patch-clamp electrophysiology. In order to increase throughput, the use of multi-electrode arrays and multiple pipettes is being explored. This would increase throughput by as much as two orders of magnitude. In addition, groups such as Bhargava *et al* have worked to increase the success rate of obtaining whole-cell patch recordings through the use of a smart patch-clamping technique which combines patch clamp electrophysiology and scanning ion conductance microscopy to scan the cell surface and generate a topographic image before recording. This allows for microdomains and, consequently, a spatial functional map of surface ion channels [34]. They have also looked into expanded the size of the probe after surface mapping to increase the likelihood of capturing channels at those locations; they found a substantially greater yield in functional data on membrane features by increasing patch pipette size [35]. This has yielded a success rate of about 70%, a dramatic increase from 40% [350]. If this method can be incorporated in vivo, patch clamp electrophysiology will become a viable method for reconstructing neural networks.

High-Throughput Animal-Behavior Experiments: In the ideal situation, high-throughput animal behavior experiments may be used to model complete human neural networks. This would allow for extended

observation and the option to work with more invasive technology. There is currently very little work being done in using these experiments for a complete model, but there has been extensive work in using animal models for the visual system. Mark Schnitzer's massively parallel two-photon imaging of the fruit fly allowed for as many as one hundred flies to be recorded at one time [52, 334]. Advances are also being made in using mice to model the human visual system; there is evidence of invariant object recognition in mice [481], as well as multifeatural shape processing [7] and transformation-tolerant object recognition. The bottleneck is not the extent to which we can put animal experiments in parallel; it is in establishing parallels to investigate. At this pace, there will be significant advances in mapping the visual system in the next two to five years using animal experimentation. A complete neural circuitry, however, may not be accomplished using this method for the next five to ten years.

Chapter 3

Psychiatry

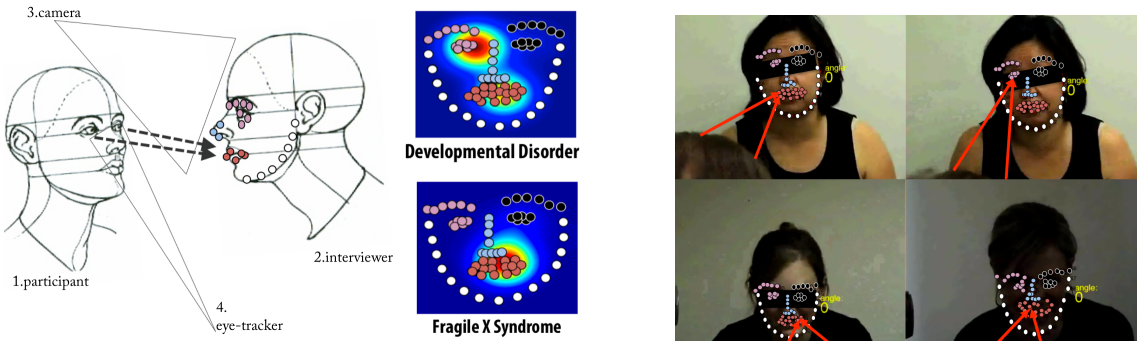
3.1 Vision-Based Classification of Developmental Disorders using Eye Movements

Autism Spectrum Disorders (ASD) is an important developmental disorder with both increasing prevalence and substantial social impact. Significant effort is spent on early diagnosis, which is critical for proper treatment. In addition, ASD is also a highly heterogeneous disorder, making diagnosis especially problematic. Today, identification of ASD requires a set of cognitive tests and hours of clinical evaluations that involve extensively testing participants and observing their behavioral patterns (e.g. their social engagement with others). Computer-assisted technologies to identify ASD are thus an important goal, potentially decreasing diagnostic costs and increasing standardization.

In this work, we focus on Fragile-X-Syndrome (FXS). FXS is the most common known genetic cause of autism [158], affecting approximately 100,000 people in the United States. Individuals with FXS exhibit a set of developmental and cognitive deficits including impairments in executive functioning, visual memory and perception, social avoidance, communication impairments and repetitive behaviors [407]. In particular, as in ASD more generally, eye-gaze avoidance during social interactions with others is a salient behavioral feature of individuals with FXS. FXS is an important case study for ASD because it can be diagnosed easily as a single-gene mutation. For our purposes, the focus on FXS means that ground-truth diagnoses are available and heterogeneity of symptoms in the affected group is reduced.

Maintaining appropriate social gaze is critical for language development, emotion recognition, social engagement, and general learning through shared attention [98]. Previous studies [220, 145] suggest that gaze fluctuations play an important role in the characterization of individuals in the autism spectrum. In this work, we study the underlying patterns of visual fixations during dyadic interactions. In particular we use those patterns to characterize different developmental disorders.

We address two problems. The first challenge is to build new features to characterize fine behaviors of



(a) (b)

Figure 3.1: (a) We study social interactions between a participant with a mental impairment and an interviewer, using multi-modal data from a remote eye-tracker and camera. The goal of the system is to achieve fine-grained classification of developmental disorders using this data. (b) A frame from videos showing the participant's view (participant's head is visible in the bottom of the frame). Eye-movements were tracked with a remote eye-tracker and mapped into the coordinate space of this video.

participants with developmental disorders. We do this by exploiting computer vision and multi-modal data to capture detailed visual fixations during dyadic interactions. The second challenge is to use these features to build a system capable of discriminating between developmental disorders. The remainder of the paper is structured as follows: In section 2, we discuss prior work. In section 3, we describe the raw data: its collection and the sensors used. In section 4, we describe the built features and analyze them. In section 5, describe our classification techniques. In section 5, we describe the experiments and results. In section 6 we discuss the results.

3.1.1 Previous Work

Pioneering work by Rehg et al. [343] shows the potential of using coarse gaze information to measure relevant behavior in children with ASD. However, this work does not address the issue of fine-grained classification between ASD and other disorders in an automated way. Our work thus extends this work to develop a means for disorder classification via multi-modal data. In addition, some previous efforts in the classification of developmental disorders such as epilepsy and schizophrenia have relied on using electroencephalogram (EEG) recordings [235]. These methods are accurate, but they require long recording times; in addition, the use of EEG probes positioned over a participant's scalp and face can limit applicability to developmental populations. Meanwhile, eye-tracking has long been used to study autism [43, 163], but we are not aware of an automated system for inter-disorder assessment using eye-tracking such as the one proposed here.

3.1.2 Dataset

Our dataset consists of 70 videos of an clinician interviewing a participant, overlaid with the participant's point of gaze (as measure by a remote eye-tracker), first reported in [159].

The participants were diagnosed with either an idiopathic developmental disorder (DD) or Fragile-X-Syndrome (FXS). DD presents similar autistic symptoms to FXS, but does not have FXS or any other known genetic syndrome.There are known gender-related behavioral differences between FXS participants, so we further subdivided this group by gender into males (FXS-M) and females (FXS-F). There were no gender-related behavioral differences in the DD group, and genetic testing confirmed that DD participants did not have FXS.

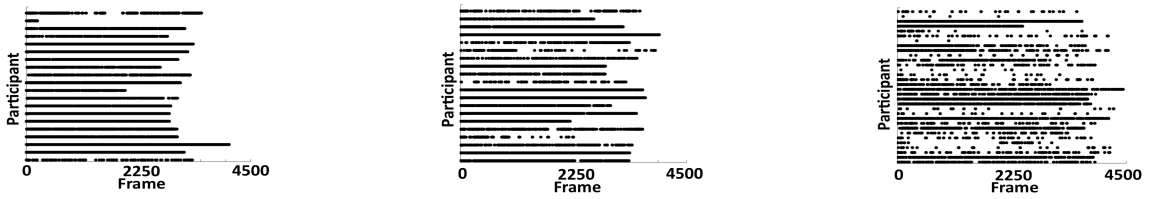
Participants were between 12 and 28 years old, with 51 FXS participants (32 male, 19 female) and 19 DD participants. The two groups were well-matched on chronological and developmental age, and had similar mean scores on the Vineland Adaptive Behavior Scales (VABS), a well-established measure of developmental functioning. The average score was 58.5 ($SD = 23.47$) for individuals with FXS and 57.7 ($SD = 16.78$) for controls, indicating that the level of cognitive functioning in both groups was 2 – 3 SDs below the typical mean.

Participants were each interviewed by a clinically-trained experimenter. In our setup the camera was placed behind the patient and facing the interviewer. Figure 3.1 depicts the configuration of the interview, and of the physical environment. Eye-movements were recorded using a Tobii X120 remote corneal reflection eye-tracker, with time-synchronized input from the scene camera. The eye-tracker was spatially calibrated to the remote camera via the patient looking at a known set of locations prior to the interview.

3.1.3 Visual Fixation Features

A goal of our work is to design features that simultaneously provide insight into these disorders and allow for accurate classification between them. These features are the building blocks of our system, and the key challenge is engineering them to properly distill the most meaningful parts out of the raw eye-tracker and video footage. We capture the participant's point of gaze and its distribution over the interviewer's face, 5 times per second during the whole interview. There are 6 relevant regions of interest: *nose, left eye, right eye, mouth, jaw, outside face*. The precise detection of these fine-grained features enables us to study small changes in participants' fixations at scale.

For each video frame, we detected a set of 69 landmarks on the interviewer's face using a part-based model [479]. Figure 3.1 shows examples of landmark detections. In total, we processed 14,414,790 landmarks. We computed 59K, 56K and 156K frames for DD, FXS-Female, and FXS-Male groups respectively. We evaluated a sample of 1K randomly selected frames, out of which only a single frame was incorrectly annotated. We mapped the eye-tracking coordinates to the facial landmark coordinates with a linear transformation. Our features take the label of the cluster (e.g. jaw) holding the closest landmark to the participant point of gaze. We next present some descriptive analyses of these data.

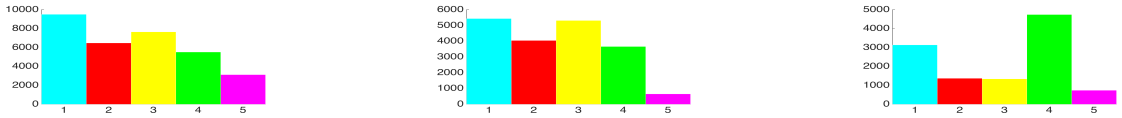


(a) DD (b) FXS-F (c) FXS-M

Figure 3.2: Temporal analysis of attention to face. X axis represents time in frames (in increments of 0.2 seconds). Y axis represents each participant. Black dot represent time points when the participant was looking at the interviewer's face. White space signifies that they were not.

Feature granularity. We want to analyze the relevance of our fine grained attention features. Participants—especially those with FXS—spent only a fraction of the time looking at the interviewer's face. Analyzing the time series data of when individuals are glancing at the face of their interviewer (see Figure 3.2), we observe high inter-group participant's variance. For example, most of FXS-F individual sequences could be easily confused with the other groups.

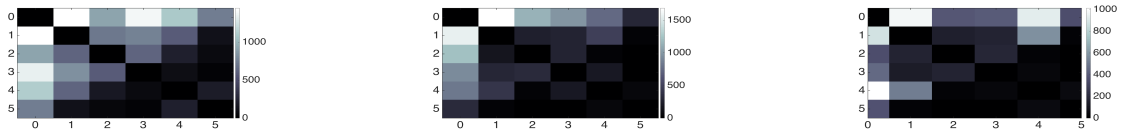
Clinicians often express the opinion that the *distribution* of fixations, not just the sheer lack of face fixations—seem related to the general autism phenotype [220, 200]. This opinion is supported by the distributions in Figure 3.3: DD and FXS-F are quite similar, whereas FXS-M is distinct. FXS-M focuses primarily on mouth (4) and nose (1) areas.



(a) DD (b) FXS-F (c) FXS-M

Figure 3.3: Histograms of visual fixation for the various disorders. X-axis represents fixations, from left to right: nose (1), eye-left (2), eye-right (3), mouth (4), and jaw (5). The histograms are computed with the data of all participants. The non-face fixation is removed for visualization convenience.

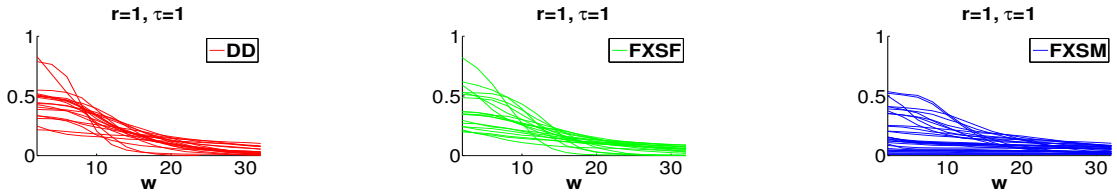
Attentional transitions. In addition to the distribution of fixations, clinicians also believe that the *sequence* of fixations describe underlying behavior. In particular, FXS participants often glance to the face quickly and then look away, or scan between non-eye regions. Figure 3.4 shows region-to-region transitions in a heatmap. There is a marked difference between the different disorders: Individuals with DD make more transitions, while those with FXS exhibit significantly less—congruent with the clinical intuition. The transitions between facial regions better identify the three groups than the transitions from non-face to face regions. FXS-M participants tend to swap their gaze quite frequently between mouth and nose, while the other two do not. DD participants exhibit much more movement between facial regions, without any clear preference. FXS-F patterns resemble DD, though the pattern is less pronounced.



(a) DD (b) FXS-F (c) FXS-M

Figure 3.4: Matrix of attentional transitions for each disorder. Each square $[ij]$ represents the aggregated number of times participants of each group transitioned attention from state i to state j. The axes represent the different states: non-face (0), nose (1), eye-left (2), eye-right (3), mouth (4), and jaw (5).

Approximate Entropy. We next estimate Approximate Entropy ($ApEn$) analysis to provide a measure of how predictable a sequence is [347] . A lower entropy value indicates a higher degree of regularity in the signal. For each group (DD, FXS-Female, FXS-Male), we selected 15 random participants sequences. We compute $ApEn$ by varying w (sliding window length). Figure 3.5 depicts this analysis. We can see that there is great variance amongst individuals of each population, many sharing similar entropy with participants of other groups. The high variability of the data sequences makes them harder to classify.



(a) DD (b) FXS-F (c) FXS-M

Figure 3.5: (a) - (c) Analysis of the $ApEn$ of the data per individual varying the window length parameter w. Y-axis is $ApEn$ and X-axis varies w. Each line represents one participant's data. We observe great variance among individuals.

3.1.4 Classifiers

The goal of this work is to create an end-to-end system for classification of developmental disorders from raw visual information. So far we have introduced features that capture social attentional information and analyzed their temporal structure. We next need to construct methods capable of utilizing these features to predict the specific disorder of the patient.

Model (RNN). The Recurrent Neural Network (RNN) is a generalization of feedforward neural networks to sequences. Our deep learning model is an adaptation of the attention-enhanced RNN architecture proposed by Hinton et al. [439] (LSTM+A). The model has produced impressive results in other domains such as language modeling and speech processing. Our feature sequences fit this data profile. In addition, an encoder-decoder RNN architecture allows us to experiment with sequences of varying lengths in a cost-effective manner. Our

actual models differ from LSTM+A in two ways. First, we have replaced the LSTM cells with GRU cells [86], which are are memory-efficient and could provide a better fit to our data [204]. Second, our decoder produces a single output value (i.e. class). The decoder is a single-unit multi-layered RNN (without unfolding) and with a soft-max output layer. Conceptually it could be seen as a many-to-one RNN, but we present it as a configuration of [439] given its proximity and our adoption of the attention mechanism.

For our experiments, we used 3 RNN configurations: RNN_128: 3 layers of 128 units; RNN_256: 3 layers of 256 units; RNN_512: 2 layers of 512 units. These parameters were selected considering our GPU memory allocation limitation.
We trained our models for a total of 1000 epochs. We used batches of sequences, SGD with momentum and max gradient normalization (0.5).

Other Classifiers. We also trained shallow baseline classifiers. We engineer a convolutional neural network approach (CNN) that can exploit the local-temporal relationship of our data. It is composed of one hidden layer of 6 convolutional units followed by point-wise sigmoidal nonlinearities. The feature vectors computed across the units are concatenated and fed to an output layer composed of an affinity transformation followed by another sigmoid function. We also trained support vector machines (SVMs), Naive Bayes (NB) classifiers, and Hidden Markov Models (HMMs).

3.1.5 Experiments and Results

By varying the classification methods described in Section 3.1.4 we perform a quantitative evaluation of the overall system. We assume the gender of the patient is known, and select the clinically-relevant pair-wise classification experiments DD vs FXS-F and DD vs FXS-M. For the experiments we use 32 FXS-male, 19 FXS-female and 19 DD participants. To maintain equal data distribution in training and testing we build S_{train} and and S_{test} randomly shuffling participants of each class ensuring a 50%/50% distribution of the two participant classes over the sets. At each new training/testing fold the process is repeated so that the average classification results will represent the entire set of participants. We classify the developmental disorder of the participants, given their individual time-series feature data p, to evaluate the precision of our system. For N total participants, we create an 80%/20% training/testing dataset such that no participant's data is shared between the two datasets. For each experiment, we performed 10-fold cross validation where each fold was defined by a new random 80/20 split of the participants –about 80 participant's were tested per experiment.

Metric. We consider the binary classification of an unknown participant as having DD or FXS. We adopt a voting strategy where, given a patient's data $p = [f_1, f_2, ...f_T]$, we classify all sub-sequences s of p of fixed length w using a sliding-window approach. In our experiments, w correspond to 3, 10, and 50 seconds of video footage. To predict the participant's disorder, we employ a max-voting scheme over each class. The

	window length	DD vs FXS-female (precision)	DD vs FXS-male (precision)
SVM	3	0.65	0.83
	10	0.65	0.80
	50	0.55	0.85
N.B	3	0.60	0.85
	10	0.60	0.87
	50	0.60	0.75
HMM	3	0.67	0.81
	10	0.66	0.82
	50	0.68	0.74
CNN	3	0.68	0.82
	10	0.68	0.90
	50	0.55	0.77
RNN_128	3	0.69	0.79
RNN_250	10	0.79	0.81
RNN_512	50	**0.86**	**0.91**

Table 3.1: Comparison of precision of our system against other classifiers. Columns denote pairwise classification precision of participants for DD vs FXS-female and DD vs FXS-male binary classification. Classifiers are run on 3,10, and 50 seconds time windows. We compare the system classifier, RNN to CNN, SVM, NB, and HMM algorithms.

predicted class C of the participant is given by:

$$C = \underset{c \in \{C_1, C_2\}}{\operatorname{argmax}} \sum_{\text{sub-seq. } s} \mathbf{1}(\text{Class}(s) = c) \tag{3.1}$$

Where $C_1, C_2 \in \{\text{DD}, \text{FXS-F}, \text{FXS-M}\}$, $\text{Class}(s)$ is the output of a classifier given input s. We use 10 cross validation folds to compute the average classification precision.

Results. The results are reported in Table 3.1. We find that the highest average precision is attained using the RNN.512 model with a 50 second time window. It classifies DD versus FXS-F with 0.86 precision and DD versus FXS-M with 0.91 precision. We suspect that the salient results produced by the RNN_512 are related to its high capacity and its capability of representing complex temporal structures.

3.1.6 Conclusion

We hereby demonstrate the use of computer vision and machine learning techniques in a cost-effective system for assistive diagnosis of developmental disorders that exhibit visual phenotypic expression in social interactions. Data of experimenters interviewing participants with developmental disorders was collected using video and a remote eye-tracker. We built visual features corresponding to fine grained attentional fixations, and developed classification models using these features to discern between FXS and idiopathic developmental disorder. Despite finding a high degree of variance and noise in the signals used, our high accuracies imply the existence of temporal structures in the data.

This work serves as a proof of concept of the power of modern computer vision systems in assistive development disorder diagnosis. We are able to provide a high-probability prediction about specific developmental diagnoses based on a short eye-movement recording. This system, along with similar ones, could be leveraged for remarkably faster screening of individuals. Future work will consider extending this capability to a greater range of disorders and improving the classification accuracy.

Chapter 4

Drug Screening

4.1 In-silico Labeling: Predicting fluorescent labels in unlabeled images

Microscopy is a uniquely powerful tool. It offers a way to observe cells and molecules across space and time. However, biological samples are mostly water and poorly refractile, so visualizing cellular structure is challenging. Optical and electronic techniques amplify contrast and make small signals visible to the human eye, but resolving other features requires different techniques, particularly fluorescence labeling. Fluorescence labeling with dyes or dye-conjugated antibodies provides unprecedented opportunities to reveal macromolecular structures, metabolites, and other subcellular constituents.

Yet, fluorescence labeling itself has limitations. Specificity varies, labeling is time consuming, specialized reagents are required, and some types of labeling perturb or even kill the cell. Immunocytochemistry commonly produces non-specific signals because of antibody cross-reactivity. Lastly, measuring the label requires an optical system that can reliably distinguish it from other signals in the sample.

We wondered if microscopic images of unlabeled cells contain more information than the human brain can readily comprehend and if new computational approaches could see more. With deep learning (DL), neural networks have been trained to achieve superhuman performance on specialized tasks [387, 409, 376]. Although promising, using DL to analyze microscopy images has been limited, often relying on pre-processed images [58, 477] or the imposition of special and somewhat artificial sample preparation procedures, such as the requirement for low plating density [174, 477, 430]. As such, it is unclear whether DL approaches provide a significant and broad-based advance in image analysis and extract information from unlabeled images that eludes the human eye.

Much of the focus of DL has been image classification, where a single label is predicted from a given image (e.g., predicting the label cat if the image contains a cat). Unfortunately, the task of predicting fluorescence images from transmitted light images is not well served by typical classification models [410] because

they typically contain spatial reductions which destroy fine detail. In response, researchers have developed specialized models for predicting images from images, including DeepLab [81] and U-Net [354]. However, we had limited success with these networks (Supplementary information, Fig. 4.18) and thus created a new one.

Here, we sought to determine if computers can find and predict features in unlabeled images that normally only become visible with invasive labeling. We designed a DL network and trained it on paired sets of unlabeled and labeled images. Using additional unlabeled images never seen by the network, we showed that features from unlabeled images of fixed or live cells accurately predict the location and texture of cell nuclei, the health of a cell, the type of cell in a mixture of cells, and the type of subcellular process. We also showed that the trained network exhibits transfer learning: it learned generalized features to solve new problems based on a very limited training set.

4.1.1 Results

Training and testing data sets for supervised machine learning

To learn to predict fluorescence images from transmitted light images, we created a dataset of training examples: pairs of pixel-registered transmitted light z-stack images and fluorescence images. To benefit from multi-task learning, in which a model is improved by learning several tasks, the training examples came from arbitrary experiments, with arbitrary samples, imaging modalities, and fluorescent labels (Fig. 4.1a). We chose deep neural networks as the statistical model to learn from the dataset: they can be expressive and result in systems with substantially superhuman performance (Fig. 4.1b). We trained the model by fitting parameters to the dataset to learn the correspondence rule (Fig. 4.1c). The trained model is a function mapping from the set of z-stacks of transmitted light images to the set of images of all fluorescent labels in the training set. If the hypothesis is correct, the model would take an unseen z-stack of transmitted light images (Fig. 4.1d) and generate images of corresponding fluorescent signals (Fig. 4.1e). Performance is measured by the similarity of predicted fluorescence and true images for held-out examples.

We generated training datasets (Table 4.2) from different cell types with different labels made by different laboratories. We used human motor neurons from induced pluripotent stem cells (iPSCs), primary murine cortical cultures, and a breast cancer cell line. Hoechst or DAPI were used to label cell nuclei, CellMask was used to label plasma membrane, and propidium iodide was used to label cells with compromised membranes. Some cells were immunolabeled with antibodies against the neuron-specific -tubulin III (TuJ1) protein, the Islet1 protein for identifying motor neurons, the dendrite-localized microtubule associated protein-2 (MAP2), or pan-axonal neurofilaments. Note that though no individual well was labeled with more than three markers, it is still possible for a model to learn to predict all labels.

To improve our chances of discovering correspondences, we collected images of unlabeled cells to maximize the information available to the network. Monolayer cultures are not strictly two dimensional, so any single image plane contains limited information about each cell. We thus collected sets of images (z-stacks)

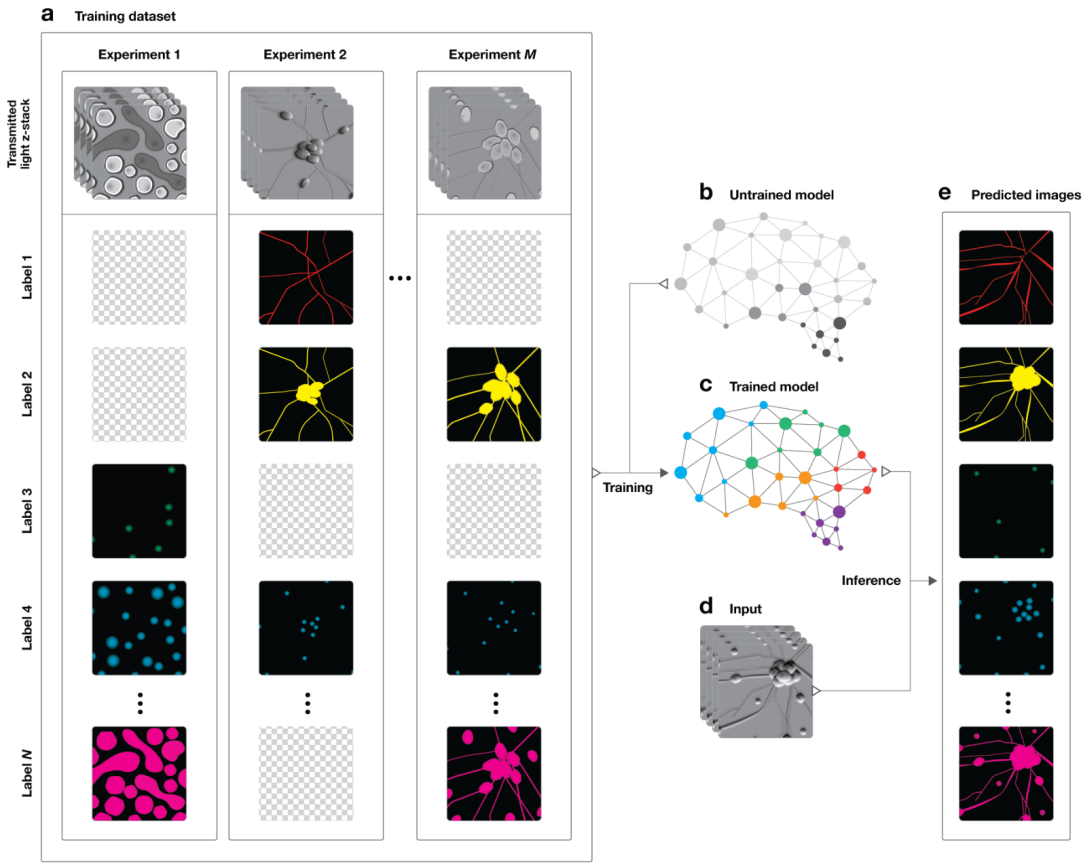


Figure 4.1: Overview of a deep learning system to train a model to make predictions of fluorescent labels from unlabeled images. (a) Dataset of training examples: pairs of transmitted light images from z-stacks of a scene with pixel-registered sets of fluorescence images of the same scene. The scenes contain varying numbers of cells; they are not crops of individual cells. The z-stacks of transmitted light microscopy images were acquired with different methods for enhancing contrast in unlabeled images. Several different fluorescent labels were used to generate fluorescence images and were varied between training examples. (b) An untrained model comprising a deep neural network with unfitted parameters was (c) trained by fitting the parameters in the untrained model to the data a. To test whether the system could make accurate predictions from novel images, a z-stack of images of a novel scene (d) were generated with one of the transmitted light microscopy methods used to produce the training data set, a. (e) The trained model, c, is used to predict fluorescence labels learned from a for each pixel in the novel images, d. The accuracy of the predictions is then evaluated by comparing them to the actual images of fluorescence labeling from d (not shown).

of the same microscope field from several planes at equidistant intervals along the z-axis and centered in the middle plane of most of the cell bodies in the field.

Translating the focal plane through the sample captures features that are in sharp focus and features out

Condition designation	Cell type	Fixed	Transmitted light	Fluorescent label #1	Fluorescent label #2	Fluorescent label #3	Training data (wells)	Testing data (wells)	Microscope field per well (μm)	Stitched image per well (pixels)‡	Laboratory
Red	Human motor neurons*	Yes	Bright field	Hoechst	Anti-TuJ1	Anti-Islet1	22	3	940 × 1300	1900 × 2600	A
Yellow	Human motor neurons*	Yes	Phase contrast	DAPI	Anti-MAP2	Anti-neurofilament	21	4	1400 × 1400	4600 × 4600	B
Green	Primary rat cortical cultures	No	Phase contrast	Hoechst	Propidium iodide	-	72	8	720 × 720	2400 × 2400	B
Blue	Primary rat cortical cultures	Yes	Phase contrast	DAPI	Anti-MAP2	Anti-neurofilament	2	1	1400 × 1400	4600 × 4600	B
Violet	Human breast cancer line	Yes	DIC	DAPI	CellMask	-	1†	1	1100 × 1100	3500 × 3500	C

*Differentiated from induced pluripotent stem cells
‡Approximate size after preprocessing.
†This condition purposely contains only a single well of training data to demonstrate that the model can learn new tasks from very little data through multitask learning.

Figure 4.2: Training data types and configurations

of focus (Fig. 4.3). Normally, out-of-focus features are undesirable, but we reasoned the implicit three-dimensional information in these blurred features could improve prediction accuracy (Supplementary information).

Next, we collected sets of images of each sample with transmitted light and fluorescence microscopy without intentionally moving the stage to minimize sample movements that might produce mis-registration of pixels between the transmitted light and fluorescence images (Fig. 4.4, Table 4.2).

Developing predictive algorithms with machine learning

With these training sets, we used supervised machine learning (ML) to determine if predictive relationships could be found between transmitted light and fluorescence images of the same cells. We used the unprocessed z-stack as input for DL algorithm development. Before applying ML, we preprocessed the images to accommodate constraints imposed by the samples, data acquisition, and ML. For example, we normalized pixel intensity distributions of the target images to make the pixel-prediction problem well defined. In addition, the image stacks were not perfectly registered along the z axis and exhibited differences in depth of field and optical sectioning. Therefore, we aimed to predict the maximum projection of the fluorescence images in the z axis rather than the entire image stack, as this projection can be easily understood visually and would be useful to predict.

We developed an ML model that is a deep neural network that performs nonlinear pixel-wise classification. It was clear that performance would benefit from a model with a multiscale input, for the same reason human eyes have foveas. We took the multiscale approach of Farabet et al.10, in which intermediate layers at multiple scales are aligned by construction, but used transposed convolutions11 to learn the resizing function rather than fixing it as in Farabet et al.. This lets the model learn the spatial interpolation rule which best fits its task. To avoid the substantial investment of manually designing and tuning a model from scratch, we chose to parameterize a space of models that could be efficiently searched over by a black box noisy function optimizer12,13 (Table 4.2).

These methods culminated in a model (methods) which achieved a lower loss on our data than other popular models while using fewer parameters (Supplementary information, Fig. 4.18) and which we judged to have satisfactory performance for several in silico labeling tasks. For each pixel of each target label image, the model produces a discrete probability distribution over 256 intensity values (corresponding to 8-bit pixels). It reads z-stacks of transmitted light images collected with bright field, phase contrast, or differential interference contrast methods and makes simultaneous predictions for every label kind that appeared in the training dataset.

The network comprises repeated modules, like the Inception network [410], but the modules and architecture differ (methods). We redesigned a version of the Inception module (Fig. 4.9), specifying widths of the layers with fewer parameters, which made it easier to search the architecture space. We used VALID convolutions (Supplementary Table 4.5) to make the network approximately position-independent, which improves scalability and correctness and removes boundary effects. The network has a multiscale input to make predictions based on a large local context (Supplementary Figs. 4.9 , 4.10, 4.13). Multiple scales are brought into geometric alignment at the midpoint of the network through the architecture rather than learned parameters. This reduces the number of variables in the model, making it easier to fit with less data.

We implemented the model in TensorFlow and trained it using the Adam optimizer [218] with asynchronous stochastic gradient descent. We optimized the hyperparameters of the DL network, such as the relative layer widths and nonlinearities, using Google Hypertune. Hypertune uses Gaussian processes for hyperparameter space modeling12 and a bandit formulation for experiment selection in a style similar to the GP-BUCB algorithm13. Hyperparameters were optimized by cross-validating on the training set; the test set was only used for final evaluation. Network predictions from transmitted light images

We asked whether we could train a network to predict the labeling of cell nuclei with Hoechst or DAPI in transmitted light images of fixed and live cells. With our trained model (Table 4.2), we made predictions of nuclear labels from images withheld from the network during the training process (Fig. 4.6). Qualitatively, the true and predicted nuclear labels looked nearly identical, and the models few mistakes appeared to be edge cases (e.g., cell-like debris lacking DNA). We created scatter plots of true versus predicted pixel intensities and quantified the correlation. Pearson ρ values 0.87 or higher indicated that the model accurately predicted the extent and level of labeling and that the predicted pixel intensities accurately reflect the true intensities, at least on a per-pixel basis. Thus, the model learned features that generalized given that these predictions were

made using different cell types and image acquisition methods.

To assess the utility of the per-pixel predictions, we gave a team of biologists real and predicted nuclear label images and asked them to annotate the images with the locations of the cell centers. With annotations on real images as ground truth, we used the methodology of Coelho et al. [93] to classify the network's errors into four categories: (Fig. 4.6b). Under conditions where cellular debris was high (e.g., Condition Yellow) or distortions in image quality evident (e.g., Condition Green), the model's precision and recall drops to the mid-90s. In other cases, the model was nearly perfect, even with dense cell clumps (e.g., Condition Blue). Network predictions of cell viability

To determine whether transmitted light images contain sufficient information to predict whether a cell is alive or dead, we trained the model with images of live cells treated with propidium iodide, a dye that preferentially labels dead cells, and made predictions from withheld images of live cells (Fig. 4.7a). The model was remarkably accurate, though not as much as it was for nuclear prediction. For example, it correctly guessed that an entity (Fig. 4.7a, second magnified outset) is actually DNA-free cell debris and not a proper cell and picked out a single dead cell in a mass of live cells (third outset). To get a quantitative grasp of the model's behavior, we created scatter plots and calculated linear fits (Fig. 4.7b). The Pearson ρ value of 0.85 for propidium iodide indicated a strong linear relationship between the true and predicted labels.

To understand the model's ability to recognize cell death and how it compared to a trained biologist, we had the real and predicted propidium iodide-labeled images annotated, following the same method as for the nuclear labels (Fig. 4.7c). A subset of the discrepancies between the two annotations in which a biologist inspecting the phase contrast images determined that an added error is a correct prediction of DNA-free cell debris was reclassified into a new category (Online Methods, Fig. 4.12). The model has an empirical precision and recall of 98% at 97%, with a 1% chance that two dead cells will be predicted to be one dead cell.

Network predictions of cell type and subcellular process type

We tested the models ability to predict which cells were neurons in mixed cultures of cells containing neurons, astrocytes, and immature dividing cells. Four biologists independently annotated real and predicted TuJ1 labeling, an indication that the cell is a neuron. We compared the annotations of each biologist (Fig. 4.8) and assessed variability among biologists by conducting pairwise comparisons of their annotations on the real labels only. With TuJ1 labels for the Condition Red culture, the performance of biologists annotating whether an object is a neuron was highly variable, consistent with the prevailing view that determining cell type based on human judgment is difficult. These measurements show humans disagree on whether an object is a neuron 10% of the time, and 2% of the time they disagree on whether an object is one cell or several cells. When a biologist was presented with true and predicted labels of the same sample, 11-15% of the time the type of cell is scored differently from one occasion to the next, and 2-3% of the time the number of cells is scored differently. Thus, the level of inconsistency introduced by using the predicted labels instead of the true labels is comparable to the level of inconsistency between biologists evaluating the same true labels.

Given the success of the model in predicting whether a cell is a neuron, we wondered whether it also could accurately predict whether a neurite extending from a cell was an axon or a dendrite. The task suffers from a global coherence problem (Supplementary information), and it was also unclear to us a priori whether transmitted light images contained enough information to distinguish dendrites from axons. Surprisingly, the final model could predict independent dendrite and axon labels (Fig. 4.14). It does well in predicting dendrites in conditions of low (Condition Yellow) and high (Condition Blue) plating density, whereas the axon predictions are much better under conditions of low plating densities (Condition Yellow).

Adapting the generic learned model to new datasets: Transfer learning

Does the network require large training data sets to learn to predict new things? Or does the generic model represented by a trained network enable it to learn new relationships in different data sets more quickly or with less training data than an untrained network? To address these questions, we attempted to use transfer learning and the trained network to learn a label from a single well. To further emulate the experience of a new practitioner adapting this technique to their problem, we chose data using a new label from a different cell type, imaged with a different transmitted light technology, produced by a laboratory other than those that provided the previous training data. In Condition Violet, differential interference contrast imaging was used to collect transmitted light data from unlabeled cancer cells, and CellMask, a membrane label, was used to collect foreground data (Table 4.2). With only the 1100-m square center of the one training well, the model learned to predict cell foreground with a Pearson score of 0.95 (Fig. 4.15). Though that metric was computed on a single test well, the test images of the well contain 12 million pixels each and hundreds of cells. These findings demonstrate that the network we trained can share learned features across tasks, a property called transfer learning. This suggests that the generic model represented by the trained network could continue to improve its performance with additional training examples, and increase the ability and speed with which it learns to perform new tasks.

4.1.2 Discussion

Here we report a new approach: in silico labeling (ISL). This ML system can infer fluorescent labels from transmitted light images. The DL network we developed could be trained on unlabeled images to make accurate per pixel predictions of the location and intensity of nuclear labeling with DAPI or Hoechst dye and to indicate if cells were dead or alive by predicting propidium iodide labeling. We showed further that the network could be trained to accurately distinguish neurons from other cells in mixed cultures and to predict whether a neurite is an axon or dendrite. These predictions showed a high correlation between the location and intensity of the actual and predicted pixels. They were accurate for live cells, enabling longitudinal fluorescence-like imaging with no additional sample preparation and minimal impact to cells. Thus, we conclude that unlabeled images contain substantial information - some not readily apparent to the human eye - that can be used to train DL networks to predict labels in both live and fixed cells that normally require invasive approaches to reveal, or which cannot be revealed using current methods.

DL has been applied to achieve useful advances in basic segmentation of microscopy images, an initial step in image analysis to distinguish foreground from background [430, 354, 84, 107, 267, 461], and on segmented images of morphologically simple cells to classify cell shape [477] and predict mitotic state [174]. Long et al. [259] applied DL methods to unlabeled and unsegmented images of low-density cultures with mixtures of three cell types and trained a network to classify cell types. Initially, we tried to adapt state-of-the-art DL models, but they had severe deficiencies. An unpublished deconvolution model based on Inception [410] produced edge artifacts and poor fine-grained detail. DeepLab [81] also had poor fine-grained detail. These issues were possibly due to a combination of locality-destroying transformations (such as max pooling with stride > 1) and artifacts introduced by zero-padding in convolutions.

Our DL network comprises repeated modules, like the reported Inception network, but the modules and architecture differ in important ways (Online methods). Inspired by U-Net [354], it is constructed so that fine-grain information can flow from the input to the output without being degraded by locality-destroying transformations. It is multiscale to provide context, but uses an architecture in which the correct spatial alignment between the scales is enforced by the network architecture rather than being learned. It preserves approximate position-independence in the output through the exclusive use of VALID transformations (Supplementary Table 4.5), which eliminates boundary effects in the predicted images. Finally, it is entirely specified as the repeated application of a single parameterized module, which simplifies the design space and makes it tractable to automatically search over network architectures.

We also gained insights into the strengths, limitations, and potential applications of DL for biologists. The accurate predictions at a per-pixel level indicate that direct correspondences exist between unlabeled images and at least some fluorescent labels. Moreover, the high-correlation coefficients for several labels indicate that the unlabeled images contain the information for a DL network to accurately predict the location and intensity of the fluorescent label. The fact that successful predictions were made under differing conditions suggests that the approach is robust and may have wide applications. Fluorescent labeling is time consuming and resource intensive, and the number of labels is limited by spectral overlap. ISL may offer, at negligible additional cost, a computational approach to reliably infer more labels than would be feasible to collect otherwise from an unlabeled image of a single sample. Also, because ISL works on unlabeled images of live cells, repeated inferences can be done on the same cell over time without invasive labeling.

That successful predictions could be made by a singly-trained network on data from three laboratories suggests that the training features were robust and generalizable. We showed that the trained network could learn a new fluorescent label from a very limited set of unlabeled data collected with a different microscopy method. This suggests that the trained network exhibited transfer learning. In transfer learning, the more a model has learned, the less data it needs to learn a new similar task. It applies previous lessons to new tasks. Thus, this network could improve with additional training data and might make accurate predictions on a broader set of data than we have measured.

Nevertheless, we encountered clear limitations of the current models predictive ability. With supervised ML, the quality of predictions is limited by the information contained in the input data. For example, the

model was less successful in identifying axons in high-density cultures. Although the model identified neurons in mixed cultures well, it was unsuccessful in predicting the motor neuron subtype (Fig. 4.16). The accuracy will be limited if there is little or no correspondence between pixels in the unlabeled image and those in the fluorescently labeled one, if the quality of labeling is severely affected due to contributions from non-specific binding or variability, or if the data are insufficient. We found from error analysis that the performance of the model depended on the amount of information in the unlabeled images, as measured by the number of images in the z-stack (Fig. 4.17). One challenge is the empirical quality of DL approaches. Network architecture and training approaches can be optimized to perform at impressive levels, but it can be difficult to determine general principles of how the network made or failed to make predictions that might guide future improvements. This will be an important area for future research.

4.1.3 Acknowledgements

We thank Lance Davidow for technical assistance, Mariya Barch for advice and helpful discussions about the manuscript, Marija Pavlovic for preparing the Condition Violet samples, Francesca Rapino and Max Friesen for providing additional cell types not used in this manuscript, Michelle Dimon for helpful advice, and Amy Chou, Youness Bennani-Smires, Gary Howard, and Kelley Nelson for editorial assistance, and Michael Frumkin for supporting the project. Financial support to do this work came from Google, NIH U54 HG008105 (SF), U01 MH1050135 (SF), R01 NS083390 (SF), the Taube/Koret Center for Neurodegenerative Disease Research (SF), the ALS Association NeuroCollaborative (SF) and the Michael J Fox Foundation Head Start Program (SF).

4.1.4 Methods

Differentiation of human iPSCs into motor neurons and plating in Condition Red. The human iPSC line 1016A was differentiated as described in Rigamonti et al 2016 [349]. Briefly, iPSCs were grown to near confluency in adherent culture in mTesr media (StemCell Technologies) before being dissociated to single cells using Accutase (cat# 07920, StemCell Technologies). Single cells were seeded into a spinning bioreactor (Corning, 55 rpm) at 1x106 cells/mL in mTesr with Rock Inhibitor (10 μM) and kept in 3D suspension culture for the duration of differentiation. The next day (day 1), dual SMAD inhibitors SB431542 (10μM) and LDN 193189 (1μM) were added. On day 2, the media was switched to KSR media (15% Knockout Serum Replacement, DMEM-F12, 1x Glutamax, 1x Non-Essential Amino Acids, 1x Pen/Strep, 1x beta-mercaptoethanol; all from Life Technologies) with SB and LDN. On day 3, the KSR media was supplemented with SB, LDN, retinoic acid (Sigma, 1μM), and BDNF (R&D, 10ng/mL). Beginning on day 5 and ending on day 10, the culture was transitioned to NIM media (DMEM-F12, 1x B-27, 1x N2, 1x Glutamax, 1x Non-Essential Amino Acids, 1x Pen/Strep, 0.2mM Ascorbic Acid, 0.16% D-glucose; all from Life Technologies). On day 6, dual SMAD inhibition was removed and Smoothen Agonist was added (1 μM). On day 10, DAPT was added (2.5μM).

On day 15, the motor neuron spheres were dissociated using Accutase and DNAse. To dissociate the spheres, they were allowed to settle in a 15 mL tube, the media was removed, they were washed with PBS and then approximately 2 mL of warmed Accutase (with 100 μL DNAse) was added to the settled pellet. Next, the tube containing the cells and Accutase was swirled by hand in a 37C water bath for 5 minutes. Then, the cells were gently pipetted up and down using a 5 mL serological pipette. To quench and wash, 5 ml of NIM was added and the cells were centrifuged at 800 rpm for 5 minutes. The pellet was then re-suspended in NB media (Neurobasal, 1x B-27, 1x N2, 1x Glutamax, 1x Non-Essential Amino Acids, 1x Pen/Strep, 0.2 mM Ascorbic Acid, 0.16% D-glucose, 10 ng/mL BDNF, 10 ng/mL GDNF, 10 ng/mL CTNF) and passed through a 40 μm filter. The filter was then washed with an additional 3 mL of NB media and the cells were counted using a BioRad automated cell counter.

For plating, the Greiner μclear 96-well plate was coated overnight at 37C with 2.5 μg/mL laminin and 25 μg/mL poly-ornithine in water. The next day, the plate was washed with DPBS twice. The dissociated motor neurons were plated at 65,000 cells per well in 200 μL of NB media and grown at 37C with 5% CO2 for 48 hours to allow processes to form.

Differentiation of human iPSCs into motor neurons and plating in Condition Yellow. The human iPSC line KW-4, graciously provided by the Yamanaka lab, was differentiated to motor neurons via a modified version of the protocol in Burkhardt et al. [59]. Briefly, iPSCs were grown to confluency on Matrigel, followed by neural induction via dual SMAD inhibition (1.5 μM Dorsomorphine + 10 μM SB431542) and WNT activation (3 μM CHIR99021) for 3 days [111]. Motor neuron specification began at day 4 by addition of 1.5 μM retinoic acid and Sonic Hedgehog activation (200nM smoothened agonist and 1 μM purmorphamine). At day 22, cells were dissociated, split 1:2 and plated in the same medium supplemented with neurotrophic factors (2ng/mL BDNF & GDNF). At day 27, neurons were dissociated to single cells using 0.05% Trypsin and plated into a 96 well plate at various cell densities (3.7K 100K/well) for fixation and immunocytochemistry.

Culturing of primary rodent cortical neurons and plating in Condition Green and Condition Blue. Rat primary cultures of cortical neurons were dissected from rat pup cortices at embryonic days 20-21. Brain cortices were dissected in dissociation medium (DM) with kynurenic acid (1 mM final) (DM/KY). DM was made from 81.8 mM Na2SO4, 30 mM K2SO4, 5.8 mM MgCl2, 0.25 mM CaCl2, 1 mM HEPES, 20 mM glucose, 0.001% phenol red and 0.16 mM NaOH. The 10x KY solution, was made from 10 mM KY, 0.0025% phenol red, 5 mM HEPES and 100 mM MgCl2. The cortices were treated with papain (100 U, Worthington Biochemical) for 10 minutes, followed by treatment with trypsin inhibitor solution (15 mg/mL trypsin inhibitor, Sigma) for 10 minutes. Both solutions were made up in DM/KY, sterile filtered and kept in a 37C water bath. The cortices were then gently triturated to dissociate single neurons in Opti-MEM (Thermo Fisher Scientific) and glucose medium (20mM). Primary rodent cortical neurons were plated into 96 well plates at a density of 25,000 cells/mL. Two hours after plating, the plating medium was replaced with Neurobasal growth medium with 100X GlutaMAX, Pen/Strep and B27 supplement (NB medium).

Culturing human cancer cells in Condition Violet. The human breast cancer cell line MDA-MB-231

was obtained from ATCC (Catalog # HTB-26) and grown in Dulbeccos modified Eagle medium (DMEM) supplemented with 10% fetal bovine sera (FBS). 15,000 cells in 150 μL of medium were used to seed each well of a 96-well plate. Cells were grown at 37C for 2 days prior to labeling.

Fluorescent labeling in Condition Red. 96 well plates were first fixed with a final concentration of 4% PFA by adding an equal volume as already present in each well of 8% PFA to each well. The plate was fixed for 15 minutes at room temperature. Next, the plate was washed with 200 μL/well of DPBS 3 times for 5 minutes each. To permeabilize the cells, they were incubated in 0.1% Triton in DPBS for 15 minutes. Again, the cells were washed with 200 μL/well of DPBS 3 times for 5 minutes each. The cells were then blocked with 1% BSA, 5% FBS in DPBS for 1 hour at room temperature. Primary antibodies were then added in blocking solution overnight at 4C at the following concentrations: rbIslet 1:1000 (Abcam cat#109517), msTuj1 1:1000 (Biolegend cat# 801202). The next day, cells were washed with blocking solution 3 times for 5 minutes each. Secondary antibodies, gtrb Alexa 488 and gtms Alexa 546, were used at 1:1000 in blocking buffer and incubated for 45 minutes at room temperature protected from light. Next, Hoechst was added at 1:5000 in DPBS for 15 minutes at room temperature protected from light. The cells were then washed with 200 μL/well of DPBS, 3 times for 5 minutes each protected from light. The cells were imaged in at least 200uL/well of clean DPBS to avoid evaporation during long scan times.

Fluorescent labeling in Condition Yellow. Day 27 iPSC-derived motor neurons were fixed in 4% Paraformaldehyde for 15 minutes, and washed 3x in DPBS. Neurons were blocked and permeabilized using 0.1% Triton-X, 2% FBS and 4% BSA for 1 hour at room temperature, and then stained with MAP2 (Abcam ab5392, 1:10000) and NFH (Encor RPCA-NF-H, 1:1000) at 4C overnight. Cells were then washed 3x with DPBS, and labeled with Alexa Fluor secondary antibodies (each 1:1000) for 1 hour at room temperature. Neurons were again washed 3x with DPBS, followed by nuclear labeling with 0.5 μg/mL DAPI.

Fluorescent labeling in Condition Green. Four day in vitro primary rat cortical neurons were treated with a cell viability fluorescent reagent (ReadyProbes Cell Viability (Blue/Green), Thermo Fisher Scientific). During treatment with the viability reagent, DMSO (1 in 1400) was added to a subset of the neurons to increase their risk of death. NucBlue Live reagent (dilution of 1 in 72) and NucGreen Dead (dilution of 1 in 144) were added to the neuronal media. The NucBlue Live reagent stained the nuclei of all cells while the NucGreen Dead reagent stained the nuclei of only dead cells. The cells were then imaged.

Fluorescent labeling in Condition Blue. Primary rat neurons were fixed in 96 well plates by adding 50 μL of 4% paraformaldehyde (PFA) with 4% sucrose to each well for 10 minutes at room temperature. PFA was removed and cells were washed three times with 200 μL of PBS. Blocking solution (0.1% Triton-x-100, 2% FBS, 4% BSA, in PBS) was added for 1 hour at room temperature. Blocking solution was removed and primary antibodies MAP2 (Abcam ab5392, 1:10000) and Anti-Neurofilament SMI-312 (BioLegend 837901, 1:500) were then added in blocking solution overnight at 4C. The next day, cells were washed with 100 μL of PBS 3 times. Cells were then treated with Alexa Fluor secondary antibodies at 1:1000 in blocking solution for 1 hour at room temperature. Neurons were again washed 3 times with PBS, followed by nuclear labeling with 0.5μg/mL DAPI.

Fluorescent labeling in Condition Violet. Adherent MDA-MB-231 cells in wells of a 96-well plate were gently washed three times by aspirating and adding 150 μL of fresh medium to remove loosely attached cells. 150 μL of medium with 3$\times$ (0.5 μL) CellMask Deep Red membrane stain (Life Technologies, Catalog #: C10046) were added to each well for a final 1.5$\times$ final concentration and incubated for 7 minutes. Samples were washed twice with fresh medium. Then, samples were fixed by aspirating media and adding 100 μL of 4% PFA to each well, prepared previously from 16% PFA in PBS (Life Technologies, Catalog #: 28906). Samples were incubated for 15 minutes more and then washed twice with PBS. PBS was aspirated and the wells were allowed to evaporate some moisture for a few of minutes. One drop of Prolong Diamond with DAPI mounting medium (Thermo Fisher, Catalog #: P36962) was added to each of the fixed wells and the plate was gently agitated to allow the mounting medium to spread evenly. Samples were placed in the refrigerator and allowed to incubate for $>$ 30 minutes before imaging.

Image acquisition. Laboratory A (Condition Red) acquired images with 40$\times$ high numerical aperture (0.95) objectives using the Operetta high content imaging microscope (Perkin Elmer) running Harmony software version 3.5.2. The illumination system for fluorescence was a Cermax Xenon fiberoptic light source. The microscope acquires images with 14-bit precision CCD cameras then automatically scales the images to 16-bit. The plate used was a 96 well Greiner μclear plate. A total of 36 wells were acquired with 36 fields representing an enclosed 66 square region. For each field, 15 planes with a distance of 0.5 μm between each were acquired. Each field overlapped with adjacent fields by 34%. Four independent channels were acquired: Bright field (50ms exposure), Hoechst (300ms exposure, 360-400 Excitation; 410-480 Emission), TuJ1 (200ms exposure, 560-580 Excitation; 590-640 Emission), and Islet1 (80ms exposure, 460-490 Excitation; 500-550 Emission). A total of 77,760 images were collected.

Laboratory B (Conditions Yellow, Green, Blue) used a Nikon Ti-E with automated ASI MS-2500 stage equipped with a spinning disc confocal microscope (Yokogawa CSU-W1), phase contrast optics [133] (Nikon S Plan Fluor 40X 0.6NA) and controlled by a custom plugin for Micro-Manager 1.4.18. An Andor Zyla4.2 camera with 2048x2048 pixels, each 6.5 μm in size, was used to generate images. For each microscope field, 1326 stacks of images were collected at equidistant intervals along the z-axis and centered in the middle plane of most of cell bodies in the field. Depending on the plate conditions, the planes in the stack were 0.31.53 μm apart, and the stack of images encompassed a total span of a 3.619.8 μm along the z-axis and centered around the midpoint of the sample. 96 well plates were used (PerkinElmer CCB). Each well was imaged with a 9 to 36 tiles (33 to 66 patterns respectively) with overlap of approximately 350 pixels. 912003 images were collected. Laboratory C (Condition Violet) used a Nikon Ti-E microscope equipped with Physik Instrumente automated stage controlled by Micro-Manager 1.4.21. Images were acquired using a confocal microscope with 1 μm z-steps with a Plan Apo 40$\times$ NA 0.95 dry objective. An Andor Zyla sCMOS camera with 6.5 μm pixel size was used, generating images with 20482048 pixels. Two wells were imaged, with 16 tiles each in a 44 pattern with approximately 300 pixel overlap.

Image acquisition with overlap. All the microscopes we used have a robotic stage for translation in the x and y dimensions, and a field of view substantially smaller than the size of the well, which provided

unsatisfying spatial context. Thus, we acquired images in sets of tiles in square tiling patterns, using the microscope's stage to translate in x and / or y between successive shots in the same well. The patterns ranged from 3×3 tiles up to 6×6. In all cases, the tiles overlapped each other to enable robust visual features based stitching into larger images. The typical overlap was about 300 pixels, which we determined as the minimum overlap required for accurate and robust stitching for a representative subset of our data (Supplementary information).

High dynamic range image acquisition. To increase the range of luminance in the image plane beyond the bit depth of the camera, we collected images in bursts of four 20-ms exposures. We then added two or more images together or averaged the group of four images on a per pixel basis to resolve features closer to the noise floor. Summing allows simple creation of images with 20, 40, 60, and 80 ms exposures. These group-summed-images provide a higher dynamic range and can then be used to reconstruct the image plane with all features more clearly visible than could be seen with any one exposure. If a direct sum of all images is used, it is possible to generate an image of the acquired plane that exceeds the bit-depth of the camera. This increases the accessible information per image plane by achieving better dynamic range and adds flexibility to the analysis allowing rescaling in bit-depth as needed.

Data preparation. The image datasets must be cleaned and canonized before they can be used to train or evaluate a ML system. To that end, they are fed through a preprocessing pipeline composed of the following stages:

1. Salt and pepper noise reduction in the fluorescence images by means of a median filter. The median filter is of size 5×5 and is applied successively until convergence, which occurs within 32 iterations.

2. *Only needed for training.* Dust artifact removal from fluorescence images, in which dust artifacts are estimated and then removed from the fluorescence images.

3. Downscaling, in which images are bilinearly downscaled by a factor of two in each dimension to reduce shot noise.

4. Flat field correction, in which the spatially varying sensitivity of the microscope is estimated and removed.

5. Dust artifact removal from transmitted light images.

6. Stitching, in which tiles with overlapping borders are montaged into a larger image, further reducing noise at the intersections while making it possible to see large parts of the well in one image.

7. *Only needed for training.* z-axis maximum projection, in which the target (fluorescence) images are projected along the z-axis by taking the 90th percentile intensity as a robust estimate of the maximum. This step is necessary to make the prediction task well-defined, because some of our confocal images had insufficient voxel z size, and because we lack a mechanism for registering voxels in the z direction across all our datasets. If we had such a system we could attempt 3D (voxel) prediction, and indeed we've had some promising results, not reported here, on a small, z-registered, dataset.

8. Global intensity normalization, in which the per-image pixel intensity distributions are constrained to have a fixed mean and standard deviation. This step, which is aided by the previous stitching step, is necessary to make the ML task well defined, because our pixel intensities are not measured in comparable absolute units. Note this would not be necessary if our samples had been instrumented with standard candles (point sources of known brightness); we would like to see in-sample calibration objects become a standard part of in vitro biology.

9. *Only needed for training.* Quality control, in which low quality images are removed from the dataset. This makes ML more tractable, as otherwise the learning system would devote resources attempting to learn the unlearnable.

Dust artifact removal from fluorescence images. A subset of the fluorescence images from the Laboratory B dataset contained the same additive intensity artifact likely due to excitation light scattering from dust. The artifact was located at the same location in each image, and appeared as a sparse pattern (¡ 10% of the pixels) of overlaid grey disks around 50 microns wide. The following procedure was used to estimate the shape and intensity of this artifact, and then to subtract it from all of the images, thereby removing the artifact. Given a collection of images all containing the artifact, the mean and minimum projections were taken across the images (i.e., for each (x,y) pixel coordinate, the mean and minimum across all images was evaluated). The sensor offset, an image sensor property, was then subtracted from the mean image, and an edge-preserving smoothing, followed by a thresholding operation, was used to produce a binary mask of the artifact location. The mask is used to replace artifact pixels in the mean image with the mean value of the non-artifact pixels, after which a Gaussian blur is applied to produce an estimate of the average background. Subtracting this average background from the average image yields the final estimate of the artifact, which is then subtracted from each of the images.

Flat field correction. Flat field miscalibration can manifest as spatially-varying image brightness consistent from image to image. We assume the effect is multiplicative and slowly spatially varying. To estimate the flat field, we take a per-pixel median across a set of images assumed to have the same bright field and then blur the result using a Gaussian kernel. The kernel standard deviation in pixels is 1/16th the image height for fluorescence images, and 1/32nd the height for transmitted light images. To flat field correct a new image, we pixelwise divide it by the flat field image and then clip the result to capture most of the intensity variation.

Dust artifact removal from transmitted light images. We treat dust in transmitted light images as a quickly spatially varying multiplicative artifact. To estimate the dust field, we take a per-pixel median across a set of images assumed to have the same dust pattern. We do not blur the images. To dust correct a new image, we pixelwise divide it by the dust field image and then clip the result to capture most of the intensity variation.

Image stitching. To stitch a set of images, we first calculate approximate (x, y) offsets between neighboring tiles using normalized cross correlation. At this point, the set of offsets may not be internally consistent; there are many paths between any two images, and the accumulated offsets along two such paths may disagree. To make the offsets internally consistent and thus refine the solution, we use a spring system

formulation and find the minimum energy configuration. In other words, for measured offsets $o_{ij} \in \Re^2$ we find the tile locations $I_i \in \Re^2$ which minimize $\sum_{i,j} ||I_i - I_j - o_{ij}||_2$. With the set of refined (x,y) offsets, we then alpha composite the tiles into a shared canvas.

Global intensity normalization. We globally affine normalize transmitted light pixel intensities to have mean 0.5 and standard deviation 0.125. We globally affine normalize fluorescence pixel intensities to have mean 0.25 and standard deviation 0.125. These numbers were chosen to make the image pixel distributions mostly fit in the $[0.0, 1.0]$ pixel intensity range, for easy visualization. Previous versions of the system had used local normalization, but it wasn't found to make much of a difference in the final images, and it contained one more knob to tune (the size of the local neighborhood).

Quality control. Of the five datasets considered in this paper, six wells were removed from Condition Red for quality concerns due to an issue with the motorized stage. This yielded the 25 remaining wells listed in Table 1.

Machine learning. Our machine learning model is a deep neural network which takes sets of transmitted light images across 13 z-depths and emits a discrete probability distribution for each pixel in each corresponding fluorescence images. Each distribution is over 256-pixel intensity values, corresponding to 8-bit pixels.

The repeated module. Inspired by Inception [410], the model is constructed by repeated applying the same basic building block (Fig. 4.9). The learned parts of the module are the two convolutions:

1. The expand convolution increases the number of features associated with each (row, column) coordinate.

2. The reduce convolution reduces the number of features associated with (row, column) coordinate.

Through hyperparameter tuning, we found that the best modules have substantially wider expansion layers than reduction layers, with an optimum ratio of about five expansion features per reduction feature. We can only speculate as to why this may be the case; though, we note others have used this pattern [339]. We also note that this finding contradicts the advice of Szegedy et al [412], in which it is argued layer widths should change gradually and monotonically. The module uses residual connections inspired by He et al. [172], this being the element-wise addition at the top of the module. However, we must define an approximate identity function, because the module always changes the layer size or scale. For the in-scale configuration, we simply trim off a size 1 border in the row and column dimensions, corresponding to a VALID convolution with a kernel size of 3 and a stride of 1. For the down-scale configuration, we do the same trim, then downscale by a factor of 2 using average pooling. For the up-scale configuration, we upscale by a factor of 2 using nearest neighbor interpolation.

Macro-level architecture. The full model is a 33-module-deep neural network, composed of the module as described (Fig. 4.10). Like U-Net9, there is a direct data path between the input and output in the native scale. The model is noteworthy in a few ways:

1. It has a multiscale input, as its input is a set of five concentric squares, where the smallest square is treated at the highest spatial detail, by the all-purple tower, and the largest square is treated at the lowest spatial detail, by the tower with red nodes.

2. As in Farabet et al. [126], the activations at the top of each tower are spatially aligned, allowing us to use a simple width concatenation to merge the towers. But unlike Farabet et al., we learn the upscaling function, allowing the network to learn the interpolation rule which best fits the task.

3. The function associated with each output node is nearly the same across all output nodes. This is nearly unheard of in complex deep neural networks. The only break from this invariant is in the convolution transpose operations in the up-scale nodes. This results in a model that more closely reflects the position-independent nature of the data, and it allows us to produce model predictions with a regular 8-pixel stride and no overlap.

The module widths were set such that each module should take roughly the same number of operations to evaluate, which means that modules get wider as their row and column size decreases. This also implies that every tower in the lower network takes roughly the same amount of time to evaluate, which is desirable for avoiding stragglers.

Training loss. For each pixel in each predicted label, the model emits a discrete probability distribution over 256 discretized pixel intensity values. The model losses are calculated as the cross-entropy errors between the predicted distributions and the true discretized pixel intensity. These cross-entropy losses are scaled such that a uniform predictor will have an error of 1.0. Each loss is gated by a pixelwise mask associated with each output channel, where the mask indicates on a per-training-datum basis whether a particular label is provided. By gating the losses in this way, we can build a multihead model on a dataset created by aggregating all our datasets. The model takes any label-free modality as input and predicts all labels ever seen. The total loss is the weighted average of the gated losses. We weighted the losses so 50% of the loss was attributed to error in predicting the fluorescence labels and 50% was attributed to error in autoencoding, in which we asked the model to predict its own inputs. We found it useful to additionally task the network with autoencoding because it can help in diagnosing training pathologies.

Training. The model was implemented in TensorFlow14 and trained using 64 worker replicas and eight parameter servers. Each worker replica had access to 32 virtual CPUs and about 20 GB of RAM. Note, GPUs would have been more efficient, but we lacked easy access to a large GPU cluster. We used the Adam [218] optimizer with a learning rate of 10E-4 for 1 week, then reduced the learning rate to 10E-5 for the second and final week. Though training for 2 weeks (about 10 million steps) was necessary to get the full performance reported here, the model converges to good predictions within the first day.

Inference. The model is applied in a sliding-window fashion, so to infer a full image, the input images are broken into patches of size 250×250 with a stride of 8, the patches are fed to the network producing outputs of size 8×8, and the outputs are stitched together into the final image. Inferring all labels on a 1024×1024 image takes about 256 seconds using 32 CPUs, or about eight thousand CPU seconds, which currently costs

about $0.01 in a public cloud. The process is parallelizable, so the inference latency can be very low, in the range of seconds. We do our own inference in parallel using Flume, a Google-internal system similar to Cloud Dataflow (https://cloud.google.com/dataflow/). The model predicts a probability distribution for each output pixel, which is useful for analyzing uncertainty. To construct images we take the median of the predicted distribution for each pixel. We've also looked at the mode (too extreme) and mean (too blurry). The predicted images don't a priori have the same average brightness as the true images, so we run them through an additional global normalization step before declaring them final.

Manual identification of model errors. To evaluate a human interpretable metric of similarity between a pair of predicted and true DAPI label images, we compared manual annotations of cell positions on each label. First, a panel of three biologists viewed the true DAPI label and identified regions to be excluded where the cell density was too high to accurately determine the cell centers in the true fluorescence images, meaning we could not score predictions in those areas. Only a small fraction of cells were excluded (Fig. 4.11), and the model made plausible (though unscorable) predictions in those regions. Next, cell center coordinates were manually annotated in the remaining regions on each of the true and predicted DAPI labels. For each coordinate, a disc shape of fixed diameter approximately the size of a cell was assigned to each annotated cell center coordinate. We took the annotations on the true label to be the ground truth reference. Following Coelho et al. [93], one-directional correspondences between objects (disc shapes) in the true and predicted labels were determined by using maximum area of overlap and the errors were classified into four types: split, merged, added, and missing. Cells at the edges of the field of view were excluded from these metrics. We then take the accuracy to be the total number of objects in the true label, less the sum of the four types of errors, divided by the total number of objects in the true label (Fig. 4.11).

The dead-cell-specific label (propidium iodide) was analyzed in a similar fashion as the nuclear labels, with the following differences. Masking of high-cell-density regions and removing annotation errors at the edges was not required. We noted that the predicted dead-cell-specific label often included false positives that were not in the true label, but after closer inspection of the phase contrast images, many of these false positives were determined to be true cellular debris that perhaps did not have DNA to be marked by the true label. Hence, after the annotations on the true and predicted dead-cell-specific label were completed, a different biologist viewed the input phase contrast images and attempted to reclassify each added error (false positive) into a correct cellular debris prediction (Fig. 4.12).

Finally, the TuJ1 label also required neither masking of high-cell-density regions nor exclusion of errors at edges. Here, not only did we repeat the within-person predicted and true label comparison across four independent biologists, but we also analyzed the consistency of their annotations on the true label to establish a baseline for human agreement. Their four annotations on the true labels yielded 12 unique pairwise comparisons for evaluating human consistency (for any two annotations, taking each to be the ground truth in turn yielded two comparisons). We report the mean error rates across both these 12 comparisons and the four predicted-versus-true comparisons, as well as the unbiased sample standard deviation (Fig. 4.8).

Code reproducibility. The TensorFlow source code and all data, including training, test, and predictions,

will be made freely available upon paper acceptance.

4.1.5 Supplemental

Tiling: Optimal image stitching parameters

The ability to stitch together a montage of tiled images depended on a variety of factors, including sample sparsity, imaging modality, number of z-depths and channels, and the overlap between adjacent tiles. On the data we worked with, we determined that a 300-pixel overlap was sufficient to get robust stitching across most datasets. This was determined empirically by cropping the tiles smaller and applying the stitching algorithm until it could no longer successfully stitch together a test set of images.

Dependence of model performance on the number of images in the z-stacks

In this work, we used the full set of 13 transmitted light images in each z-stack (Supplementary Fig. 4.3). However, it wasnt clear a priori whether the model needs all 13 z-depths. To test this, for each N_z in $1, 2, ..., 13$ we trained independent models with N_z input z-depths. To specify which z-depths to provide the model, we used a fixed ordering of the z-stack images starting at the center plane where most of the cells should be in focus (z = 6 in a 0-indexed count) and expanding outward along the z axis in steps of two z-depths. For instance, with this strategy, to select three of the available 13 z-stacks, we would select z-depths 4, 6, and 8.

To measure the performance on a subset of N_z z-depths, we extracted N_z z-depths according to our fixed z-stack ordering and then trained an independent model on this image subset for four million steps. We then measured cross entropy loss for fluorescence image prediction on a validation set (Supplementary Fig. 4.17). Note, the curves in Supplementary Fig. 4.18 show losses for combined prediction of fluorescence labels and auto-encoding, which tend to be lower.

These experiments suggest that performance improves with the number of input z-depths, but that each additional image provides less benefit than the last. We do not find this surprising; each additional image provides additional information the model can learn to use, but eventually performance will saturate.

Limitations

Even with a superhuman machine learning system, in silico labeling (ISL) would not work when the transmitted light z-stack lacks the information needed to predict the labels:

1. Neurites are hard to discern in Condition Blue, so the axon prediction was not very accurate (Supplementary Fig. 4.14).
2. Nuclei are nearly invisible in Condition Violet, so the nuclear prediction was not very localized (Supplementary Fig. 4.15).
3. Motor neurons look like regular neurons, so the predicted motor neuron label (Islet1) was not very specific to motor neurons (Supplementary Fig. 4.16).

4.

Thus, all applications of ISL should be validated on a characteristic sample before being trusted on a new dataset.

Global coherence

The current model uses a cheap approximation to the correct loss function, not the correct loss itself. The final output of ISL is an image, but the loss we use is over pixels, not images. Thus, the model will attempt to predict the most likely pixels, and will make each of those predictions independently. This means that predicted images may lack global coherence; instead of getting clear structures in images, predictions may produce erroneous averages over several structures. Practically speaking, the problem is most noticeable for long thin structures like neurites and explains why theyre not always predicted as continuous shapes (Supplementary Fig. 4.14). The problem could be addressed with existing techniques from machine learning, e.g. sampling techniques [428] or adversarial models [149].

Comparison to other deep learning models

The proposed model outperformed the DeepLab [81] and U-Net [354] models on this data. To determine this, we trained those networks and the proposed model on our training data. The proposed model achieved a lower loss than U-Net, which achieved a lower loss than DeepLab (Supplementary Fig. 4.18). Early comparisons of the same kind were what drove us to develop a new architecture, rather than rely on existing architectures.

For each learning rate in [1e-4, 3e-5, 1e-5, 3e-6], each model was trained for at least 10 million steps using Adam [218], which took around 2 weeks each on a cluster of 64 machines. For each model, we selected the trained instance with the best error out of the 4 learning rates. For the proposed model, it was 3e-6. For DeepLab and U-Net it was 1e-5. These 3 trained instances had been continuously evaluated on the training and validation datasets, producing the training curves shown in the figure.

The DeepLab and U-Net implementations we used were provided by the Vale team at Google, which maintains internal implementations of common networks, and which created DeepLab. For U-Net, we used an input size of 321 and a batch size of one. The proposed model had 27 million trainable parameters, DeepLab had 80 million, and U-Net had 88 million.

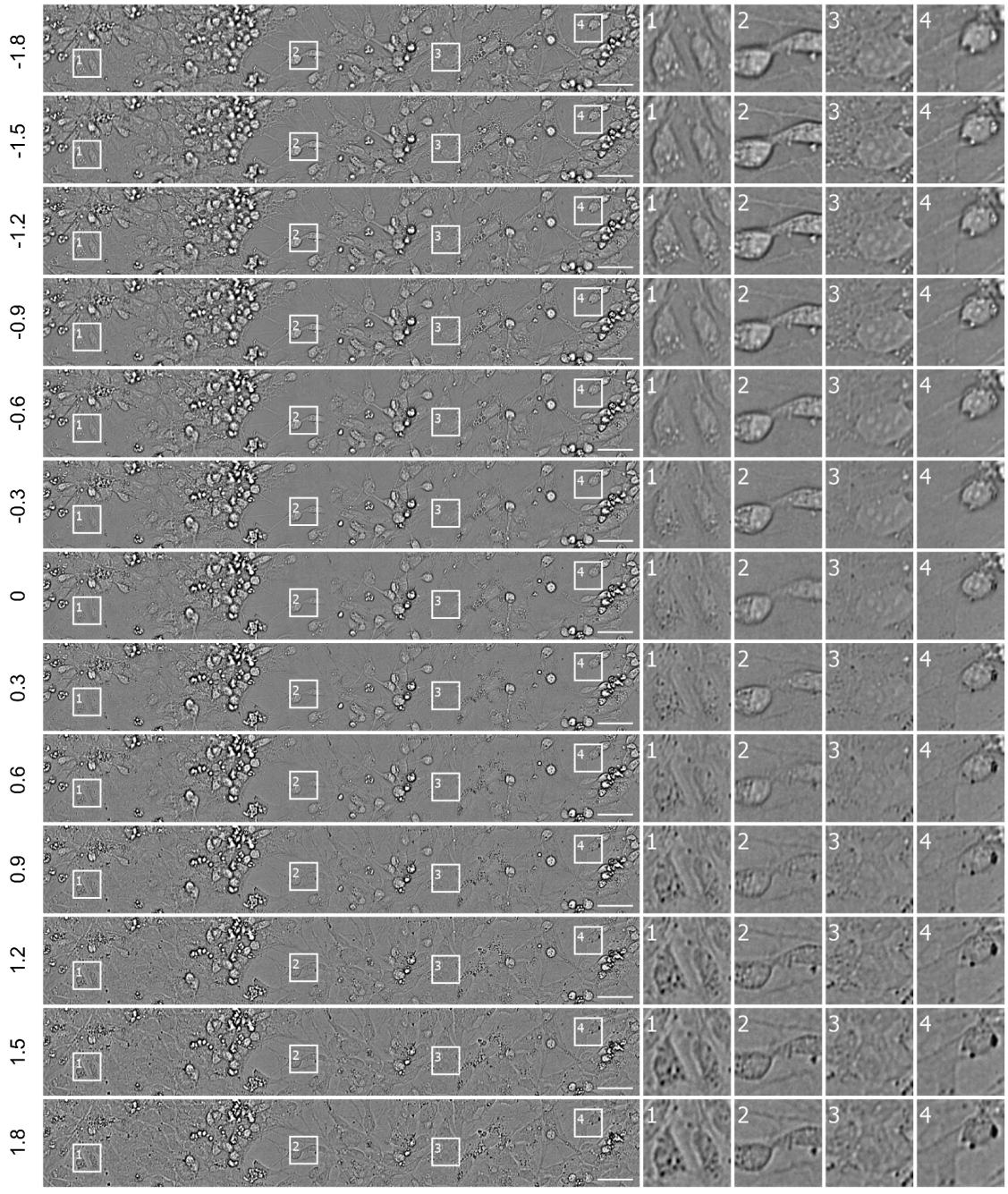


Figure 4.3: Generation of z-stacks of transmitted light images of unlabeled cells. To initially train the network and to test the predictions of the network, z-stacks of transmitted light images of a given microscope field were generated by collecting a total of 13 images: one approximately at the focal plane and an additional six images above and below that plane. In the example shown from Condition Red, the 13 images in a stack were spaced 0.3 m apart, spanning 3.6 m along the z-axis. The location of each image relative to the central plane is given in microns by the numbers to left of the images. The outsets illustrate how different planes capture different information about the sample with some planes providing greater detail about intracellular structure and others providing more information about neurites and cell morphology. Scale bars are 40 m.

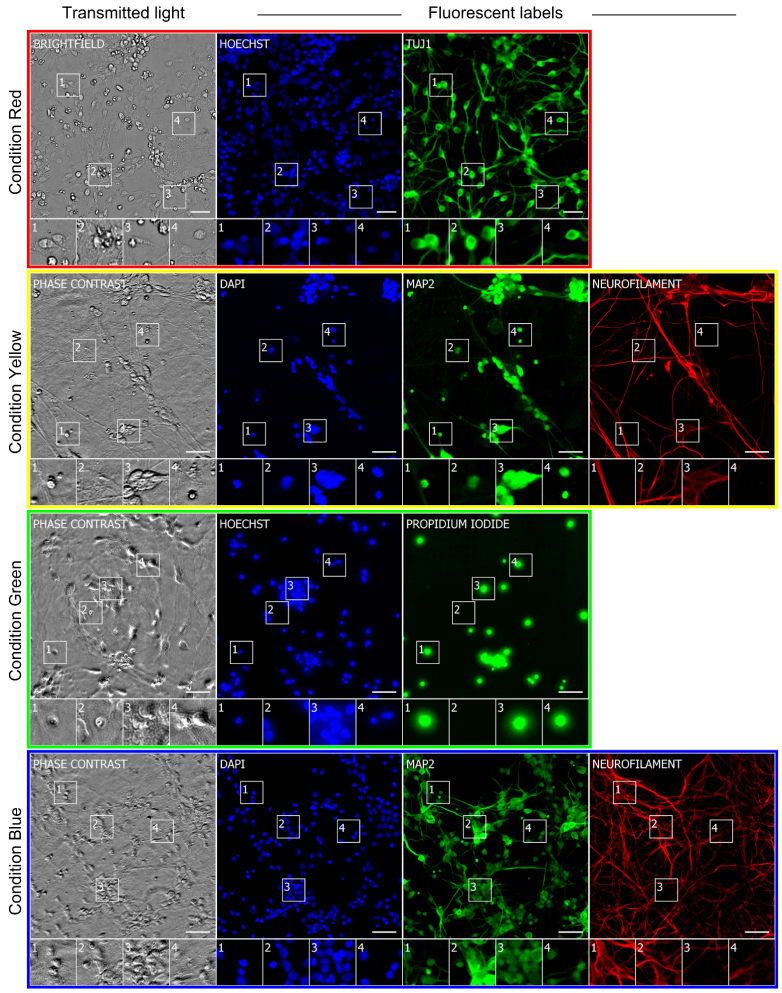


Figure 4.4: Example images of unlabeled and labeled cells used to train a deep learning network. Each row is a typical example of labeled and unlabeled images from datasets described in Table 4.2. The first column is the center image from the z-stack of unlabeled transmitted light images from which the model makes its predictions. Subsequent columns show fluorescence images of labels that the model will use to learn correspondences with the unlabeled images and eventually try to predict from unlabeled images. The numbered outsets show magnified views of subregions of images within a row. The training data are diverse: sourced from two independent laboratories using four different cell types, six fluorescent labels and both bright field and phase contrast methods to acquire transmitted light images of unlabeled cells. The scale bars are 40 m.

Concept	Description	Use in paper	References
Supervised learning	Machine learning paradigm where a model is trained with pairs of inputs and output examples ("labels"), where the goal is to be able to generalize to predict the output given new unseen inputs.	We describe an instance where a model is trained to predict labeled images from unlabeled images.	https://en.wikipedia.org/wiki/Supervised_learning
Image prediction	Machine learning applications in which the output of the statistical model is an image.	We propose an image prediction system which produces the fluorescence images which would have resulted from physical labeling, without requiring physical labeling.	-
Deep neural networks	A class of statistical model with the capacity to fit any continuous function on a bounded domain. Critical to modern superhuman AI.	We describe an instance of this class which can predict labeled images from unlabeled images.	https://en.wikipedia.org/wiki/Deep_learning
Test data set	Test data, pairs of inputs and known output examples unseen by a model, are used to assess how well a model might generalize to new data.	All results presented in the text and figures are generated on held out test images which were not seen by any model or any authors until manuscript preparation.	https://en.wikipedia.org/wiki/Test_set
Network architecture	Deep networks are composed of many operations, each of which was designed to perform particular tasks. The choice and arrangements of the operations is called the architecture. Architecture design is an unsolved problem, somewhere between art and science.	We propose a particular architecture, which we found to be satisfactory for the labeling task. First we describe a new operation (the module), and then an architecture composed of the module.	-
Convolution	An operation which applies the same function to every receptive field in a network layer. Efficient to compute and critical for image, video, and audio processing. FULL convolutions can return tensors with the same dimensions as the input, which means they often must introduce hallucinated values at the edges. VALID convolutions often change the tensor size but do not introduce hallucinated values.	The network makes exclusive use of VALID convolutions, which complicates architecture design but reduces prediction artifacts.	https://en.wikipedia.org/wiki/Convolutional_neural_network
Automatic hyperparameter tuning	Algorithms which search design spaces in an automatic fashion. They often use Bayesian modeling to eliminate from consideration bad regions in the design space.	We used around 100 million CPU hours to automatically optimize the design of the proposed module. Note, this is a one-time cost!	https://en.wikipedia.org/wiki/Hyperparameter_optimization https://cloud.google.com/ml/docs/concepts/hyperparameter-tuning-overview
Multitask learning	Related to transfer learning, it is the idea that learned abstractions can be reused across multiple tasks. Because of this reuse, the cost of training is reduced, computationally and in data requirements.	Multitask learning allows the network to make good predictions in Condition Violet, despite the paucity of training data.	https://en.wikipedia.org/wiki/Multi-task_learning https://en.wikipedia.org/wiki/Inductive_transfer

Figure 4.5: Machine learning concepts

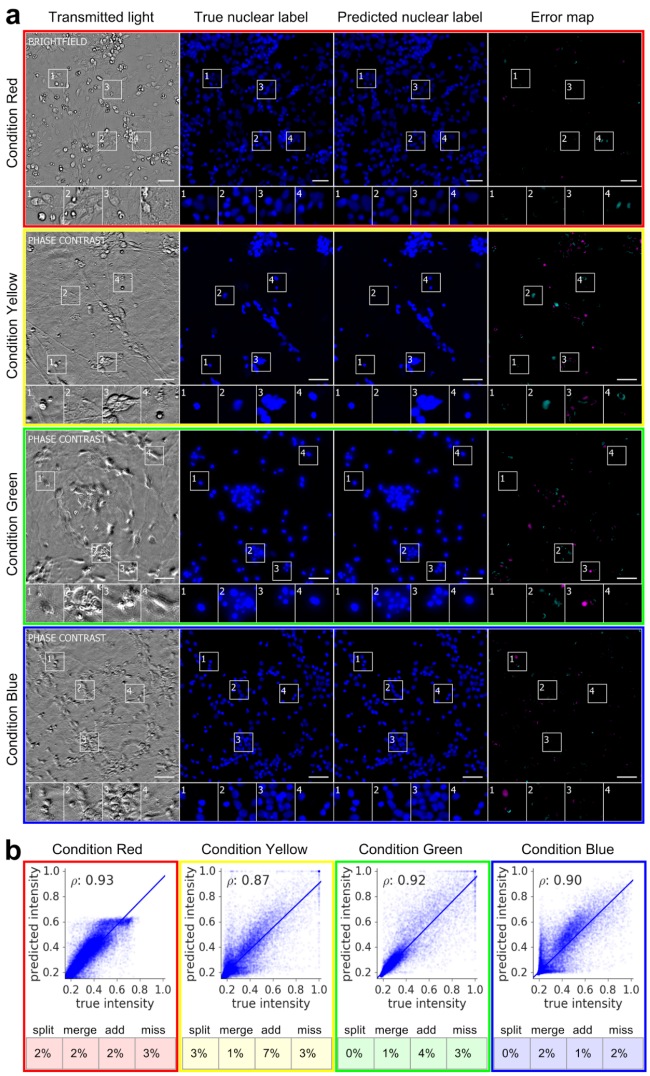


Figure 4.6: Predictions of nuclear labels (DAPI or Hoechst) from unlabeled images. (a) Upper-left-corner crops of test images from datasets in Table 4.2; please note that images in all figures are small crops from much larger images and that the crops were not cherry-picked. The first column is the center transmitted image of the z-stack of images of unlabeled cells used by the model to make its prediction. The second and third columns are the true and predicted fluorescent labels, respectively. Predicted pixels that are too bright (false positives) are magenta and those too dim (false negatives) are shown in teal. Condition Red Outset 4 and Condition Yellow Outset 2 shows false negatives. Condition Green Outset 3 and Condition Blue Outset 1 show false positives. Condition Yellow Outsets 3 and 4 and Condition Green Outset 2 show a common source of error, where the extent of the nuclear label is predicted imprecisely. Other outsets show correct predictions. Scale bars are 40 μm. (b) The scatter plots compare the true fluorescence pixel intensity to the model's predictions, with inset Pearson ρ values. The solid line is the best linear fit. See Supplementary Figure 4.19 for a detailed breakdown. Under each scatter plot is a further categorization of the errors and the percentage of time they occurred. Split is when the model mistakes one cell as two or more cells. Merged is when the model mistakes two or more cells as one. Added is when the model predicts a cell when there is none (i.e., a false positive), and missed is when the model fails to predict a cell when there is one (i.e., a false negative).

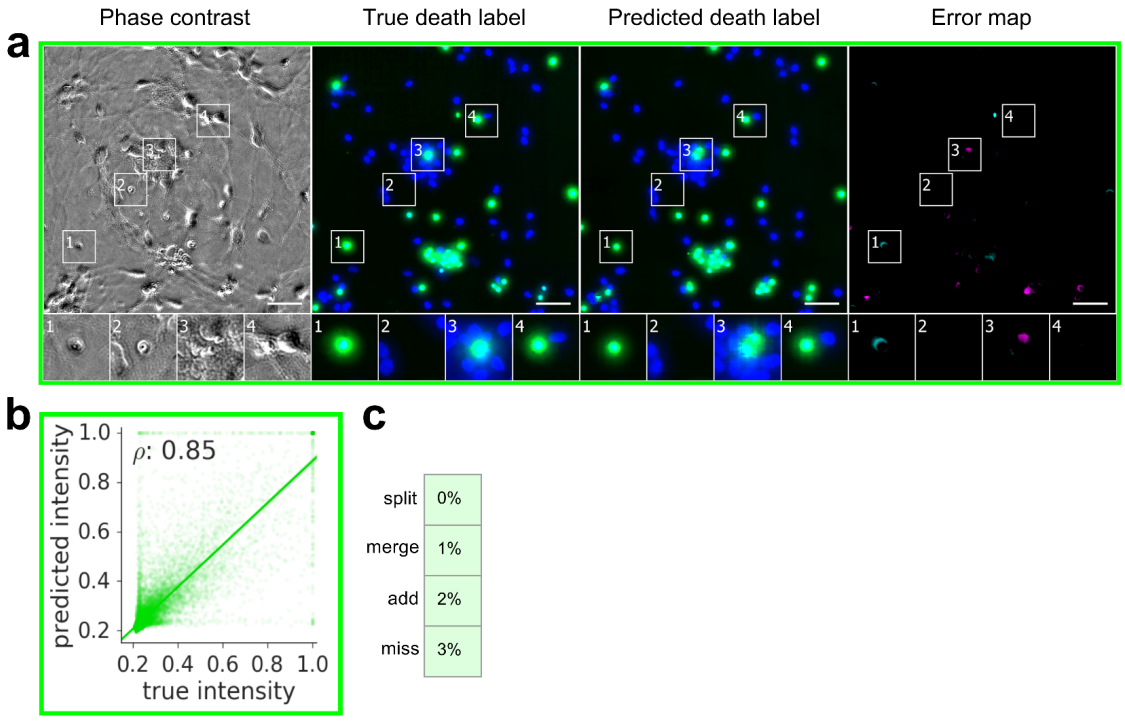


Figure 4.7: Predictions of cell viability from unlabeled live images. The trained model was tested for its ability to predict cell death, indicated by labeling with propidium iodide staining shown in green. (a) Upper-left-corner crops of cell death predictions on the datasets from Condition Green (Table 4.2). Similarly to Figure 4.6, the first column is the center phase contrast image of the z-stack of images of unlabeled cells used by the model to make its prediction. The second and third columns are the true and predicted fluorescent labels, respectively, shown in green. Predicted pixels that are too bright (false positives) are magenta and those too dim (false negatives) are shown in teal. The true (Hoechst) and predicted nuclear labels have been added in blue to the true and predicted images for visual context. Outset 1 in a shows a misprediction of the extent of a dead cell, and Outset 3 in a shows a true positive adjacent to DNA-free debris which was predicted to be propidium iodide positive. The other outsets show correct predictions. (b) The scatter plot compares the true fluorescence pixel intensity to the model's predictions, with inset Pearson values, on the full Condition Green test set. The solid line is the best linear fit. See Supplementary Figure 4.20 for a detailed breakdown. (c) A further categorization of the errors and the percentage of time they occurred. Split is when the model mistakes one cell as two or more cells. Merged is when the model mistakes two or more cells as one. Added is when the model predicts a cell when there is none (i.e. a false positive), and missed is when the model fails to predict a cell when there is one (i.e. a false negative). The scale bars are 40 μm.

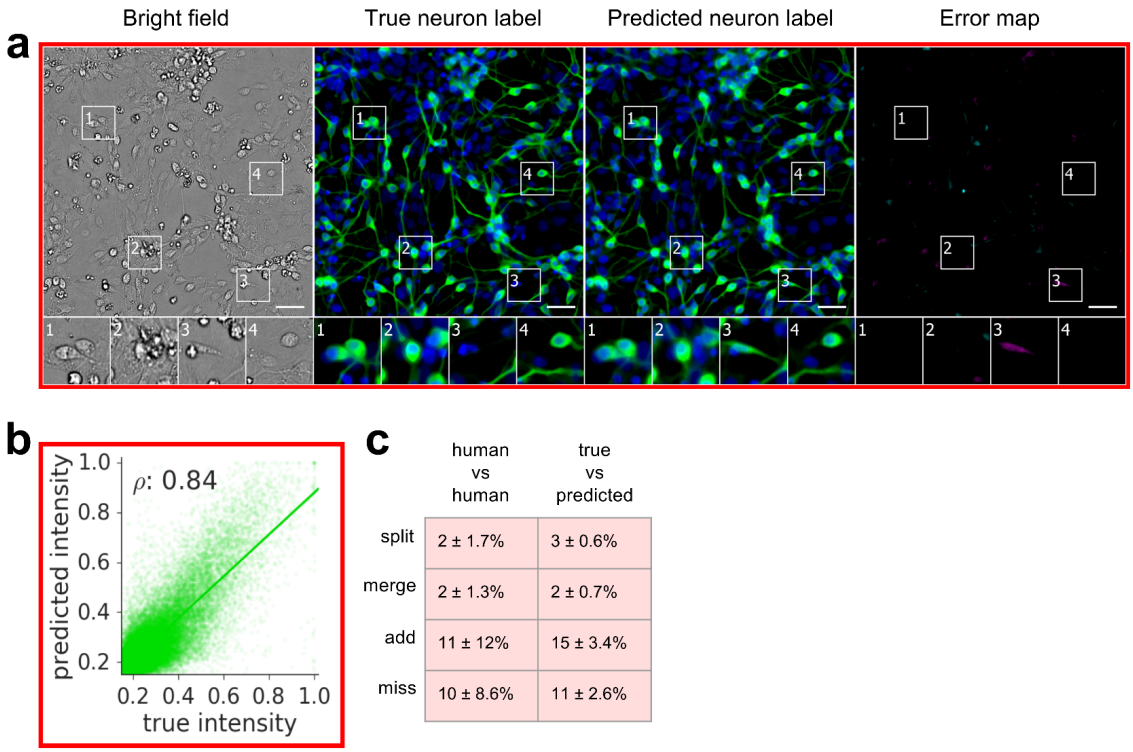


Figure 4.8: Predictions of cell type from unlabeled images. The model was tested for its ability to predict from unlabeled images which cells are neurons. The neurons come from cultures of induced pluripotent stem cells differentiated toward the motor neuron lineage but which contain mixtures of neurons, astrocytes, and immature dividing cells. (a) Upper-left-corner crops of neuron label (TuJ1) predictions, shown in green, on the Condition Red data (Table 4.2). The unlabeled image that is the basis for the prediction and the images of the true and predicted fluorescent labels are organized similarly to Figure 4.6. Predicted pixels that are too bright (false positives) are magenta and those too dim (false negatives) are shown in teal. The true and predicted nuclear (Hoechst) labels have been added in blue to the true and predicted images for visual context. Outset 3 in a shows a false positive: a cell with a neuronal morphology that was not TuJ1 positive. The other outsets show correct predictions. (b) The scatter plot compares the true fluorescence pixel intensity to the model's predictions, with inset Pearson values, on the full Condition Red test set. The solid line is the best linear fit. See Supplementary Figure 4.20 for a detailed breakdown. (c) A further categorization of the errors and the percentage of time they occurred. The error categories of split, merged, added and missed are the same as in Figure 4.6. There is an additional "human vs human" column, showing the expected disagreement between expert humans predicting which cells were neurons from the true fluorescence image, treating a random expert's annotations as ground truth. The scale bars are 40 m.

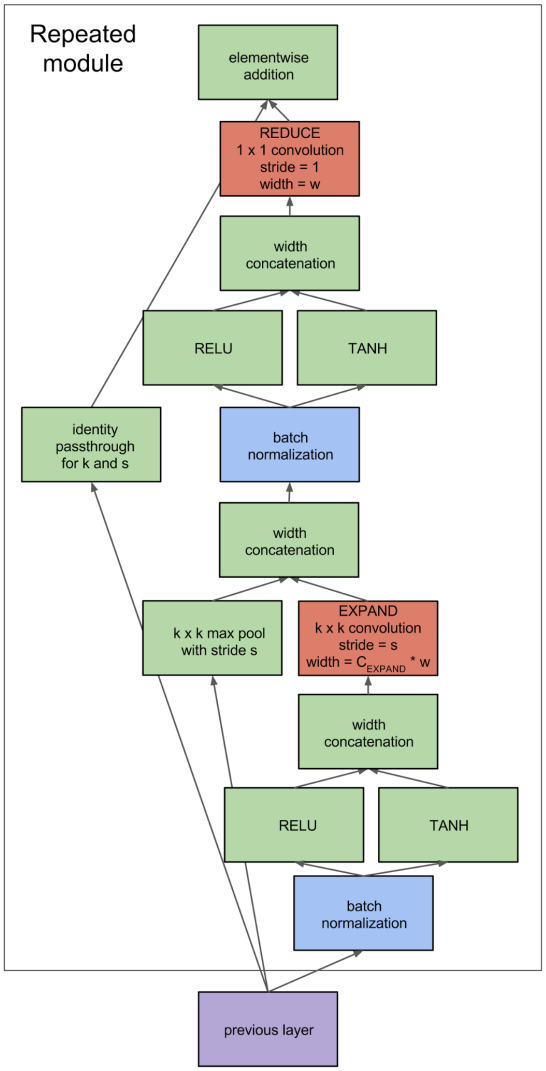


Figure 4.9: The repeated module, the basic building block of this deep network model. Data flows from the bottom to the top, along the indicated edges. Red operations contain variables to be learned, green operations have no trained variables, and blue operations are batch normalization [189]. This module is parameterized with three values: the width w, the size of the first convolution kernel k, and the stride s. CEXPAND is a constant, which we set to 5.41 after hyperparameter tuning. It is used in one of three configurations: (1) in the in-scale configuration, k = 3 and s = 1; (2) in the down-scale configuration, k = 4 and s = 2; (3) in the up-scale configuration, k = 4, s = 2, the max pool is dropped, and the expand convolution is replaced with a transposed convolution11, followed by a center crop to make the convolution transpose more space invariant. In this crop, activations within two rows or columns of the border are discarded. All convolutions and the max pooling are VALID, meaning they don't use any imputed missing activation values.

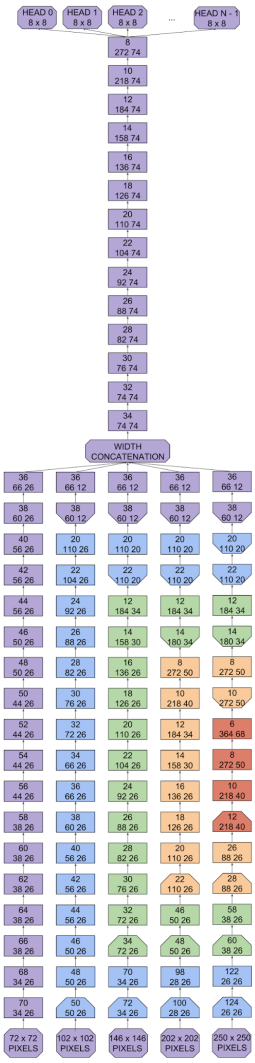


Figure 4.10: The deep neural network, the full statistical model used for label prediction. The rectangles and hexagons are the network modules: the rectangles are in-scale, the hexagons with flat bottoms are down-scale, and the hexagons with flat tops are up-scale. The octagons at the bottom are raw pixels read from the unlabeled image stack, and the octagons at the top are model heads, from which the predicted patches are derived for each fluorescent label. The colors correspond to the spatial scale of each particular module. Purple is the native scale, blue is 2× downscale, green is 4× downscale, orange is 8× downscale, and red is 16× downscale. The top number in each module is the number of rows and columns of its output layer. The bottom two numbers are the widths of the modules expansion and reduction layers, respectively. The network reads from a concentric set of five square patches, ranging in size from 72×72 pixels to 250×250 pixels, processes each one independently, merges them, does more processing, then predicts a number of 8×8 patches.

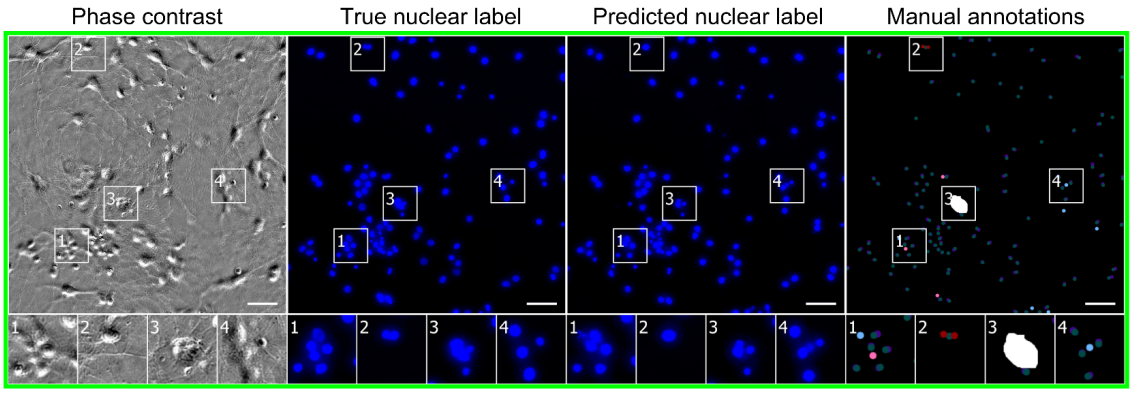


Figure 4.11: Sample manual error annotations for the nuclear label (DAPI) prediction task on the Condition Green data. The unlabeled image that is the basis for the prediction and the images of the true and predicted fluorescent labels are organized similarly to Figure 4.6, but the fourth column instead displays manual annotations. Merge errors are shown as red dots, add errors are shown as light blue dots, and miss errors are shown as pink dots. There are no split errors. All other dots indicate agreement between the true and predicted labels. Outset 1 shows an add error in the upper left, a miss error in the center, and six correct predictions. Outset 2 shows a merge error. Outset 4 shows an add error and four correct predictions. Outset 3 shows one correct prediction, and a cell clump excluded from consideration because the human annotators could not determine where the cells are in the true label image. The scale bars are 40 μm.

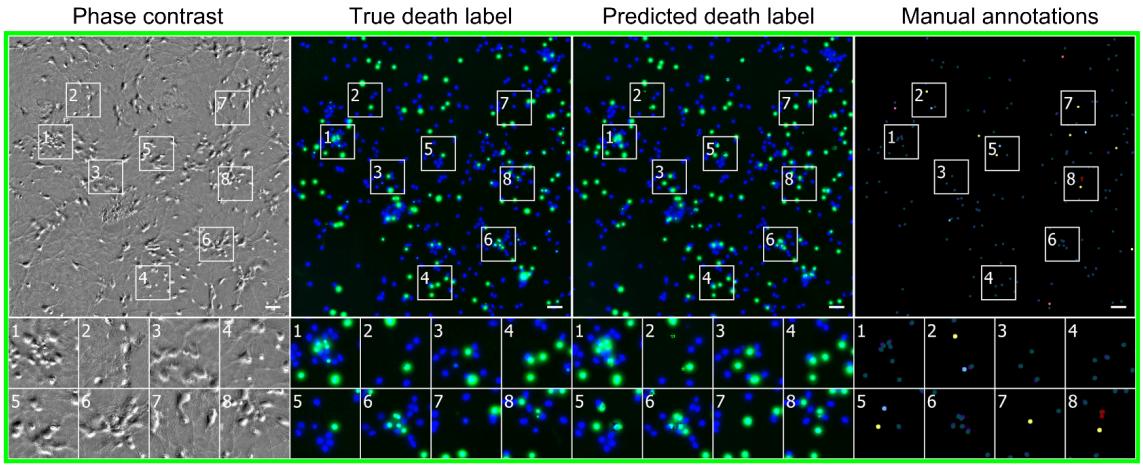


Figure 4.12: Sample manual error annotations for the cell death label (propidium iodide) prediction task on the Condition Green data. The unlabeled image that is the basis for the prediction and the images of the true and predicted fluorescent labels are organized similarly to Figures 4.6 and 4.7, but the fourth column instead displays manual annotations, and the true and predicted nuclear (DAPI) labels have been added for visual context. Merge errors are shown as red dots, add errors are shown as light blue dots, miss errors are shown as pink dots, and add errors which were reclassified as correct debris predictions are shown as yellow dots. There are no split errors. Outset 2 shows an add error at the bottom and a reclassified add error shown at top. The top error was reclassified because of the visible debris in the phase contrast image. Outset 5 shows an add error at the top and a reclassified add error at the left. Outset 7 shows a reclassified add error. Outset 8 shows a merge error at the top and a reclassified add error at the bottom. All other dots in the outsets show correct predictions. Note, the dead cell on the left in Outset 3 is slightly positive for the true death label, though it is very dim. The scale bars are 40 μm.

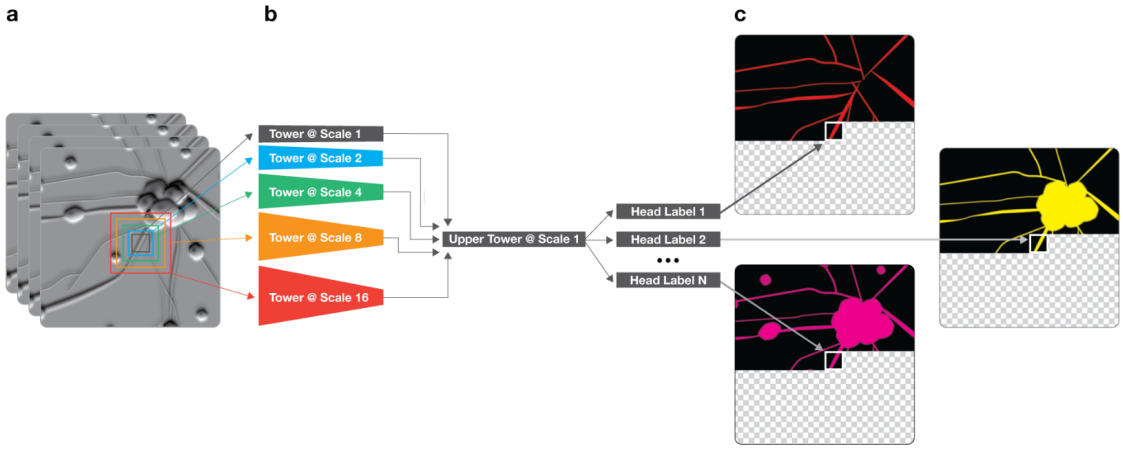


Figure 4.13: Machine learning workflow for model development. (a) Example z-stack of transmitted light images with five colored squares showing the models multiscale input. The squares range in size, increasing approximately from 72×72 pixels to 250×250 pixels, and they are all centered at the same fixation point. Each square is cropped out of the transmitted light image from the z-stack and input to the model component of the same color in b. (b) Simplified model architecture. The model is composed of six serial sub-networks (towers) and one or more pixel-distribution-valued predictors (heads). The first five towers process information at one of five spatial scales and bring the information into spatial alignment at the native spatial scale. The sixth and last tower processes the aligned information. (c) Predicted images at an intermediate stage of image prediction. The model has already predicted pixels to the upper left of its fixation point, but hasn't yet predicted pixels for the lower right part of the image. The input and output fixation points are kept in lockstep and are scanned in raster order to produce the full predicted images.

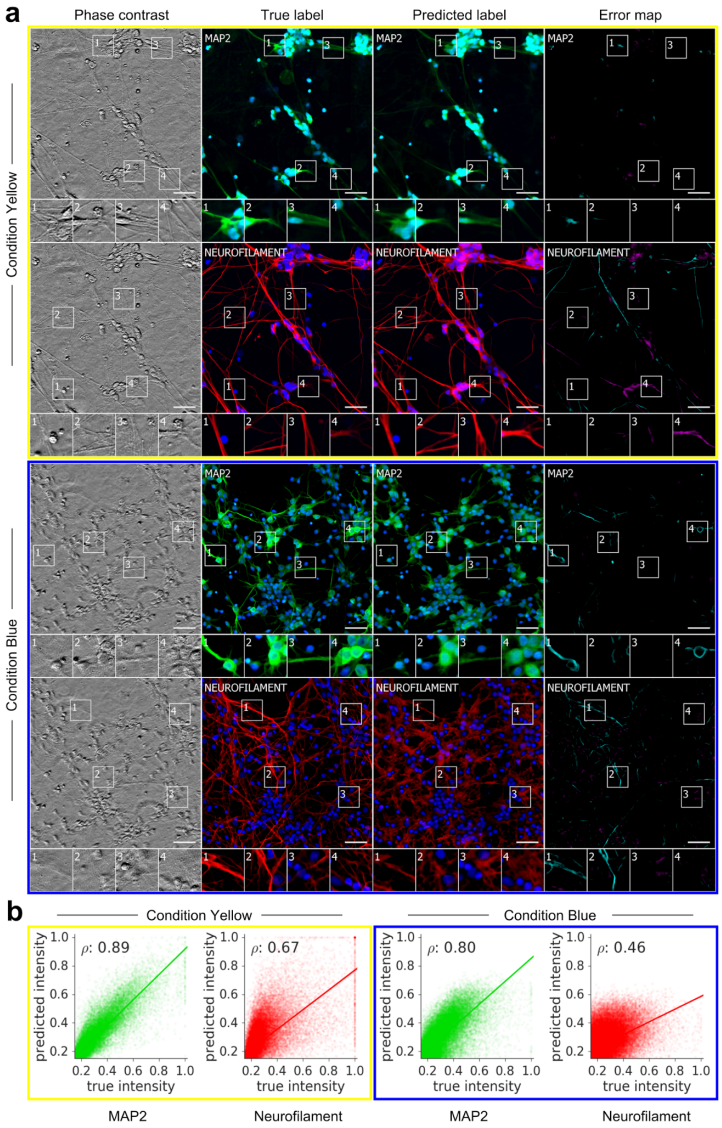


Figure 4.14: Predictions of neurite type from unlabeled images. (a) Upper-left-corner crops of dendrite (MAP2) and axon (neurofilament) label predictions on the Conditions Yellow and Blue datasets. The unlabeled image that is the basis for the prediction and the images of the true and predicted fluorescent labels are organized similarly to Figure 4.6. Predicted pixels that are too bright (false positives) are magenta and those too dim (false negatives) are shown in teal. The true and predicted nuclear (DAPI) labels have been added to the true and predicted images in blue for visual context. Outset 4 for the axon label prediction task in Condition Yellow shows a false positive, where an axon label was predicted to be brighter than it actually was. Outset 1 for the dendrite label prediction task in Condition Blue shows a false negative, where a dendrite was predicted to be an axon. Outset 4 in the same row shows an error in which the model underestimates the extent and brightness of the dendrite label. Outsets 1,2 for the axon label prediction task in Condition Blue are false negatives, where the model underestimated the brightness of the axon labels. All outsets in this row show the model does a poor job predicting fine axonal structures in Condition Blue. All other outsets show correct predictions. Scale bars are 40 m. (b) Pixel intensity scatter plots and the calculated Pearson coefficients for the correlation between the intensity of the actual label for each pixel and the predicted label. See Supplementary Figure 4.20 for a detailed breakdown.

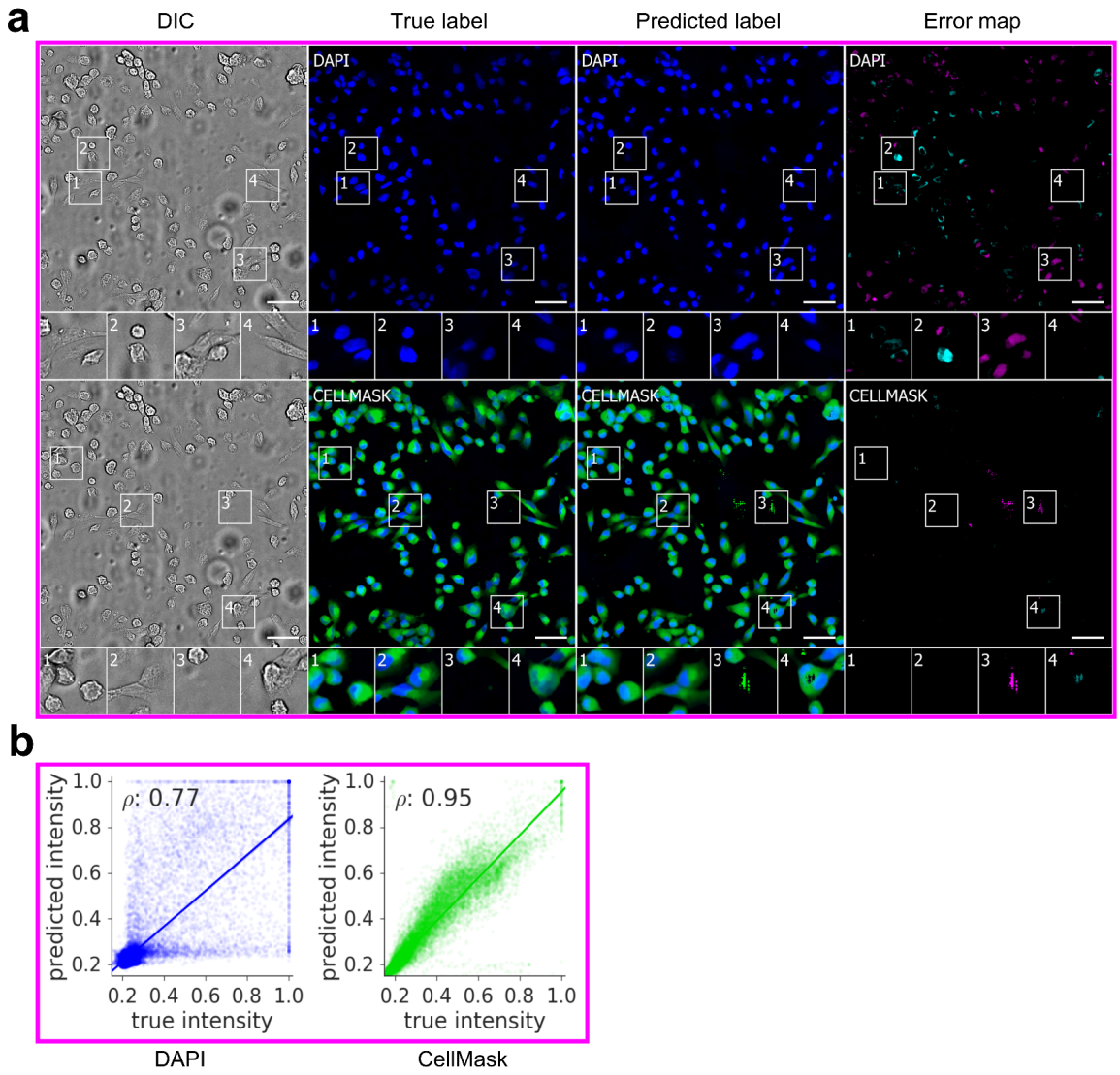


Figure 4.15: An evaluation of the ability of the trained network to exhibit transfer learning. (a) Upper-left-corner crops of nuclear (DAPI) and foreground (CellMask) label predictions on the Condition Violet dataset, representing 9% of the full image. The unlabeled image used for the prediction and the images of the true and predicted fluorescent labels are organized similarly to Figure 4.6. Predicted pixels that are too bright (false positives) are magenta and those too dim (false negatives) are shown in teal. In the second row, the true and predicted nuclear labels have been added to the true and predicted images in blue for visual context. Outset 2 for the nuclear label task shows a false negative in which the model entirely misses a nucleus below a false positive in which it overestimates the size of the nucleus. Outset 3 for the same row shows the model underestimate the sizes of nuclei. Outsets 3,4 for the foreground label task show prediction artifacts; Outset 3 is a false positive in a field that contains no cells, and Outset 4 is a false negative at a point that is clearly within a cell. All other outsets show correct predictions. The scale bars are 40 m. (b) Pixel intensity scatter plots and the calculated Pearson coefficients for the correlations between the pixel intensities of the actual and predicted label. Although very good, the predictions have visual artifacts such as clusters of very dark or very bright pixels (e.g., boxes 3 and 4, second row). These may be a product of a paucity of training data. See Supplementary Figure 4.21 for a detailed breakdown.

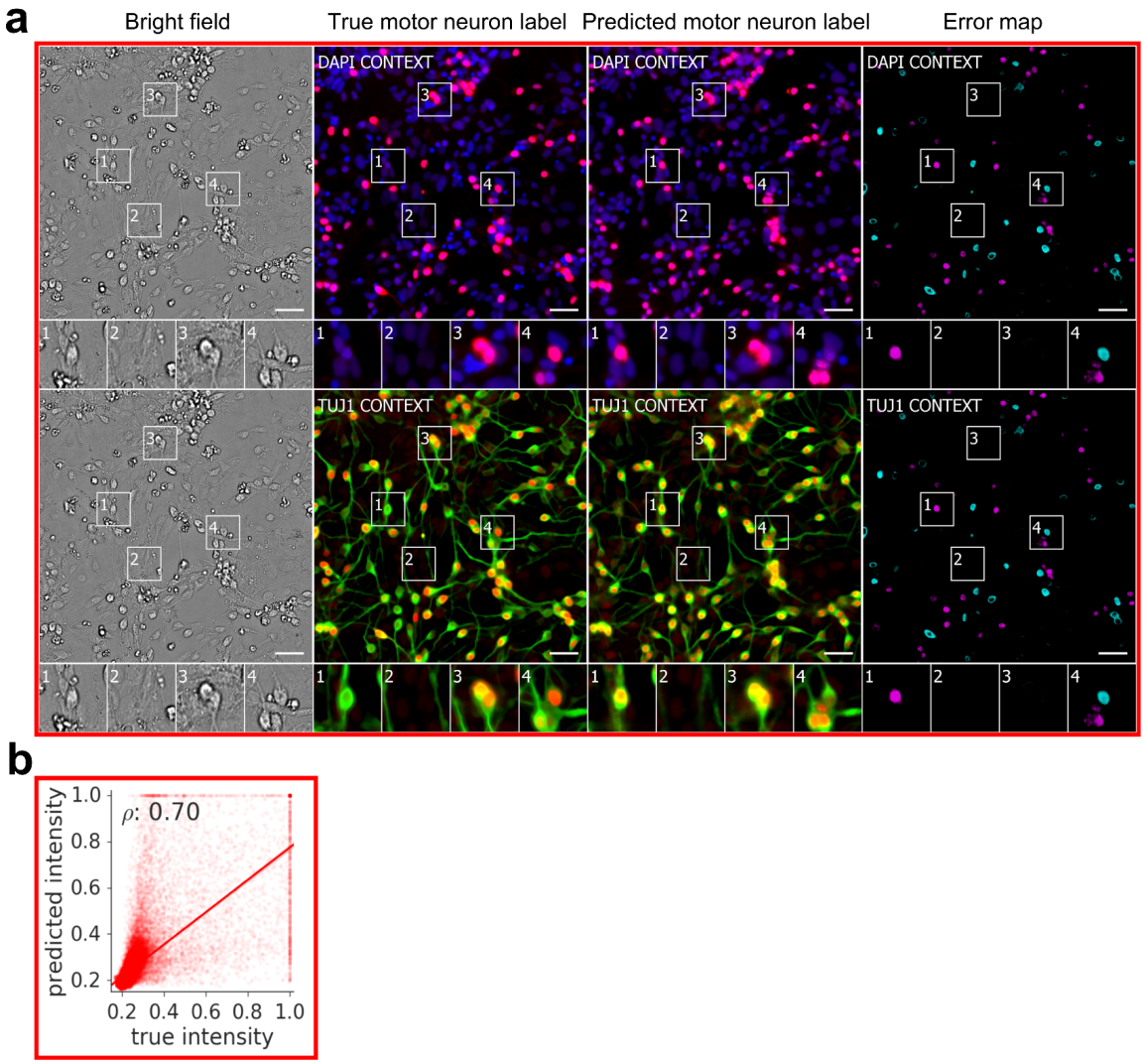


Figure 4.16: Predictions of neuron subtype from unlabeled images. (a) Upper-left-corner crops of motor neuron label (Islet1) predictions for Condition Red dataset. The unlabeled image that is the basis for the prediction and the images of the true and predicted fluorescent labels are organized similarly to Figure 4.6, but in the first row the true and predicted nuclear (DAPI) labels have been added to the true and predicted images in blue for visual context, and in the second row the true and predicted neuron (TuJ1) labels were added. Outset 1 shows a false positive, in which a neuron was wrongly predicted to be a motor neuron. Outset 4 shows a false negative above a false positive. The false negative is a motor neuron that was predicted to be a non-motor neuron, and the false positive is a non-motor neuron that was predicted to be a motor neuron. The two other outsets show correct predictions. The scale bars are 40 m. (b) Pixel intensity scatter plots and the calculated Pearson coefficients for the correlation between the intensity of the actual label for each pixel and the predicted label. See Supplementary Figure 4.21 for a detailed breakdown.

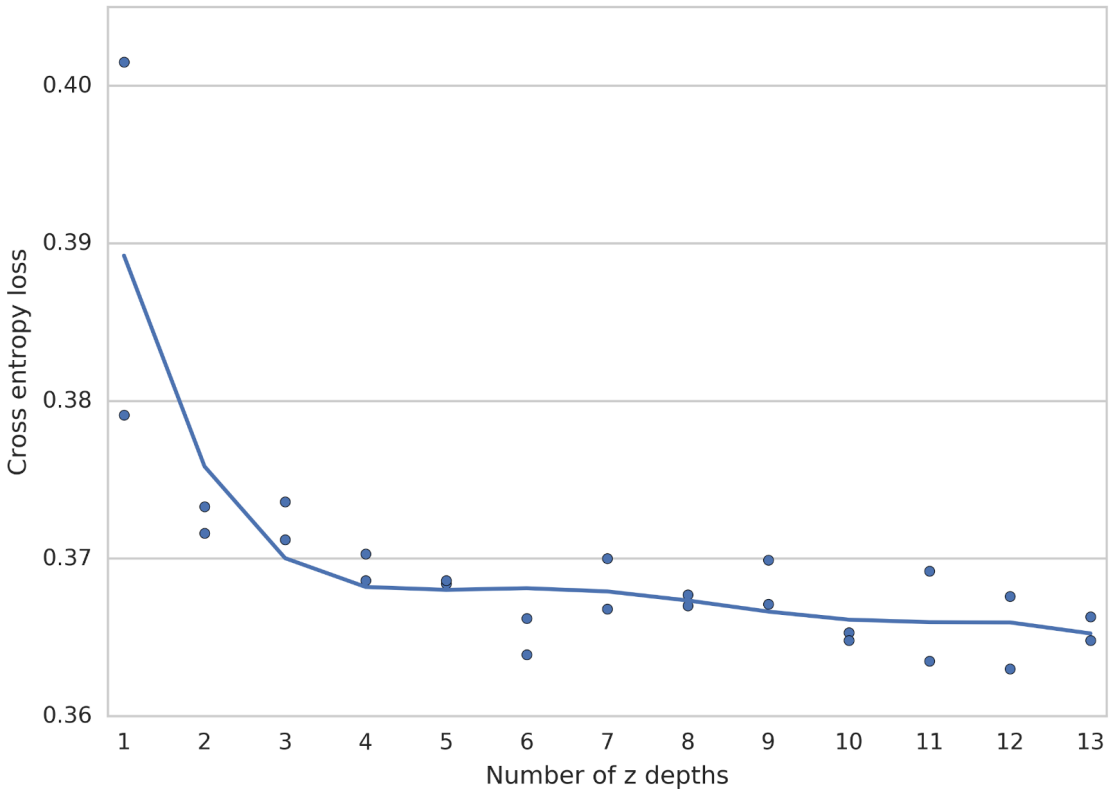


Figure 4.17: Dependence of model performance on the number of images in the transmitted light z-stack. The x-axis is the number of images in the model input. The y-axis is the cross entropy loss on fluorescence label prediction on a validation set. Each dot is the loss of a single model after training for 4 million steps with the optimal learning rate of 3e-6. Two models were trained for each configuration, yielding 26 dots. The curve is the degree 5 polynomial which best fits the data under the least squares loss. The more distinct z-depths provided to the model, the better it performs.

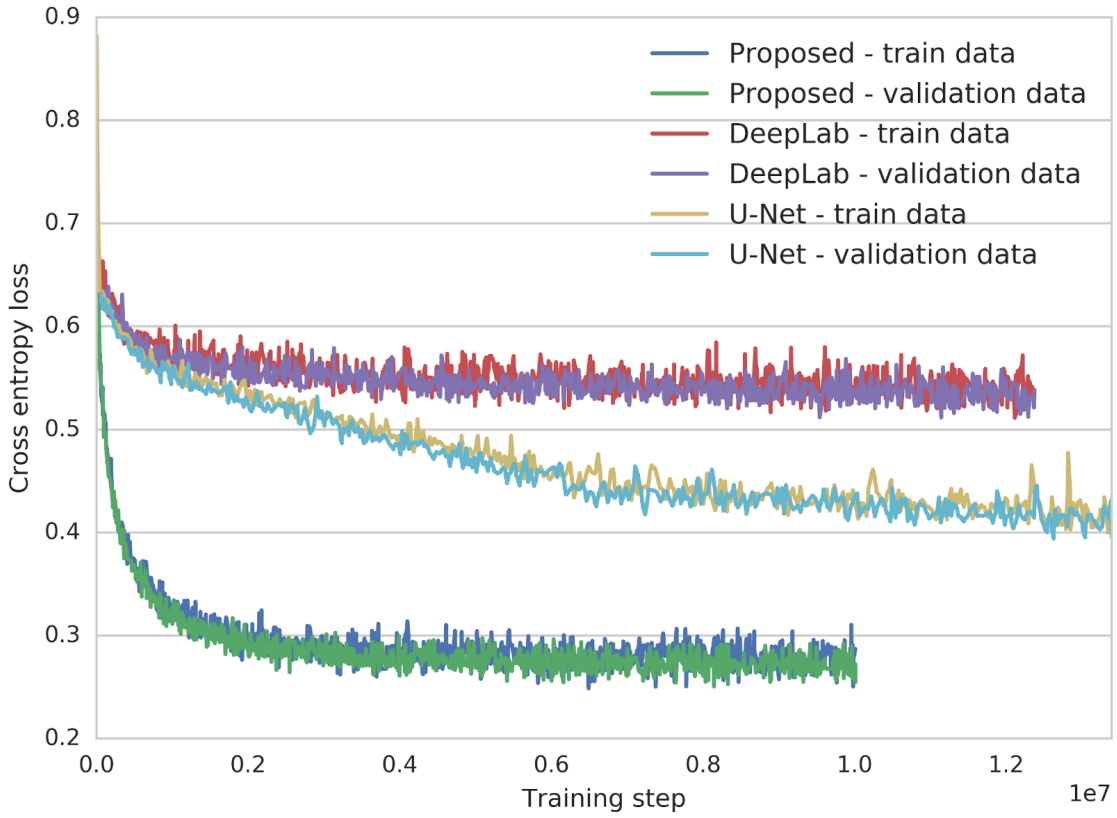


Figure 4.18: Comparison of the proposed model to DeepLab and U-Net. The curves show cross entropy loss on the training and validation data, as a function of the number of training steps. The proposed model achieved a lower loss than U-Net, which achieved a lower loss than DeepLab. All models were trained for at least 10 million steps, which took around 2 weeks per model training on a cluster of 64 machines.

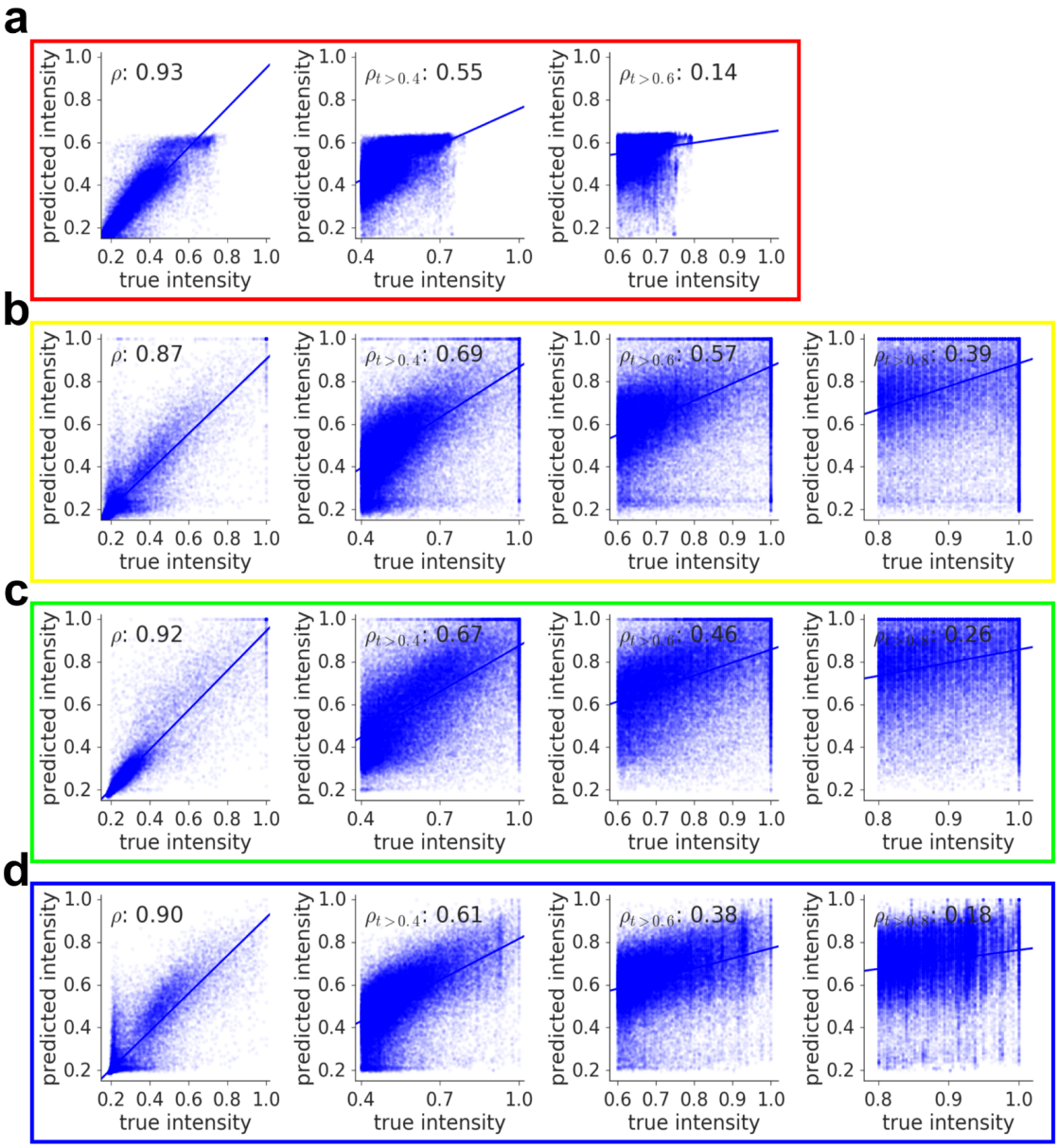


Figure 4.19: Breakdown of scatter plots from Figure 4.6. Each subfigure shows the original scatter plot, along with scatter plots restricted to true / predicted pixel pairs where the true intensity is above the indicated threshold. These additional plots help explain how well the model can predict the intensity of a pixel, given the true pixel intensity is above a certain level. (a) Condition Red. There were no true pixels with intensity $>$ 0.8 for this condition. (b) Condition Yellow. (c) Condition Green. (d) Condition Blue.

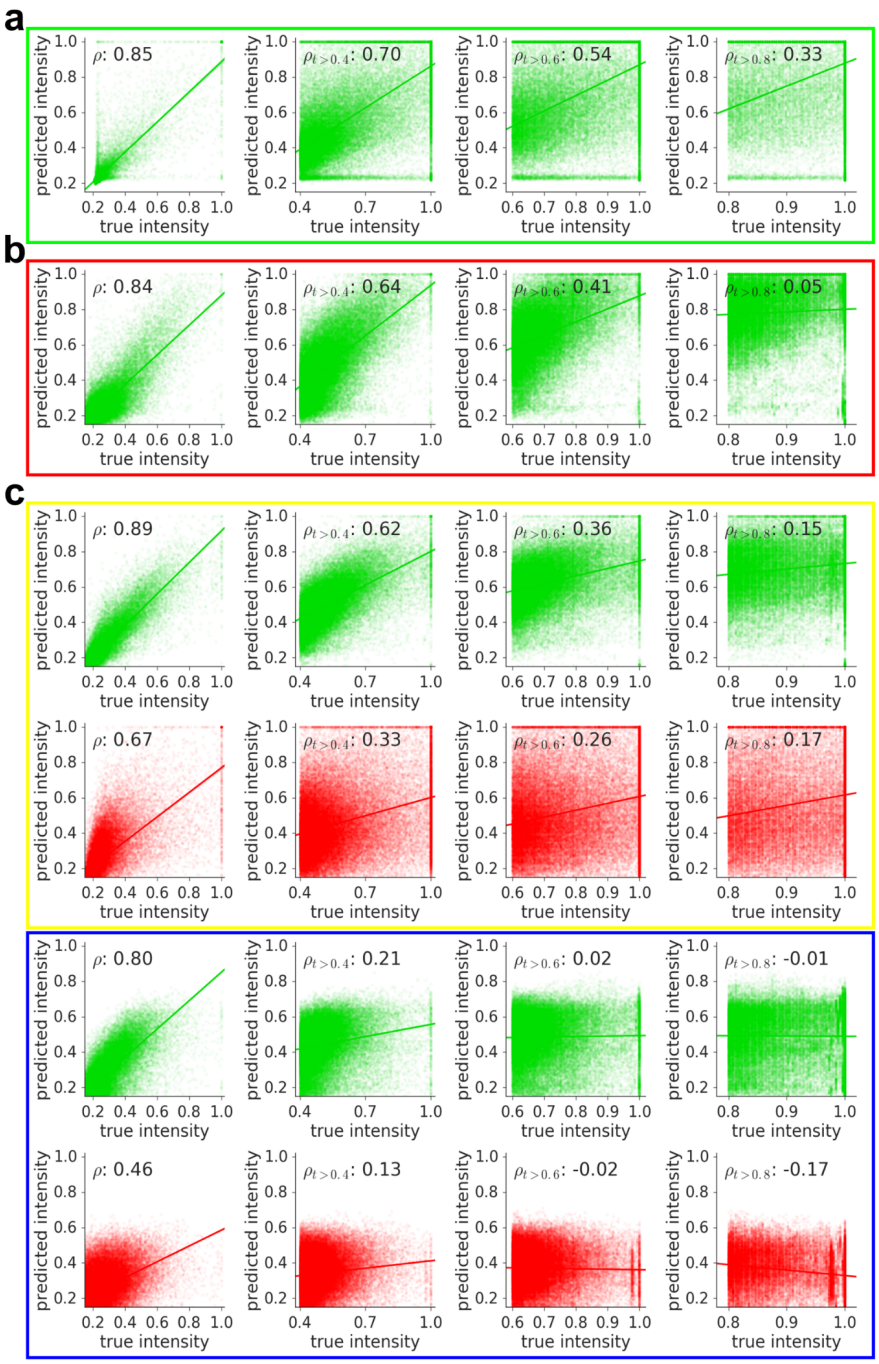


Figure 4.20: Breakdown of scatter plots from Figures 4.7 and 4.8 and Supplementary Figure 4.14 in the same style as Supplementary Figure 4.19. (a) Figure 4.7. (b) Figure 4.8. (c) Supplementary Figure 4.14. The first row is Condition Yellow, MAP2 prediction. The second row is Condition Yellow, Neurofilament prediction. The third row is Condition Blue, MAP2 prediction. The fourth row is Condition Blue, Neurofilament prediction.

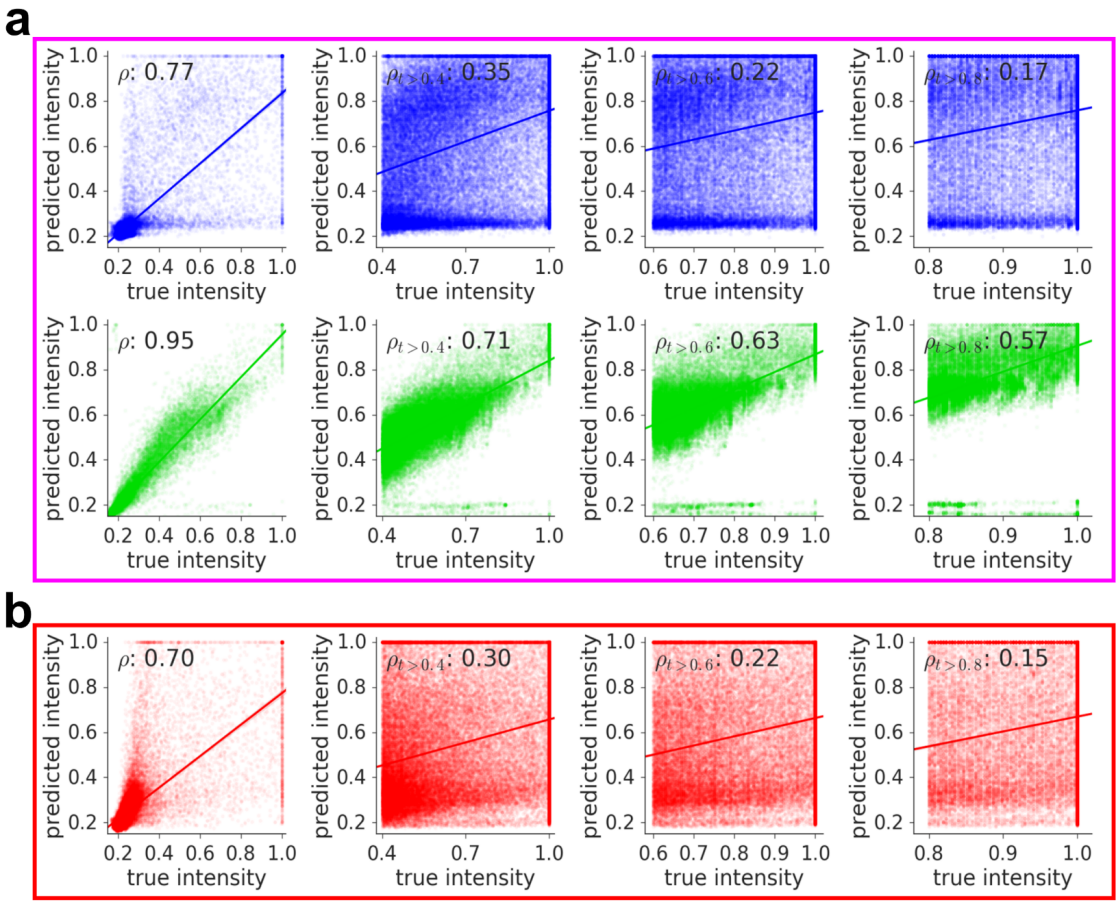


Figure 4.21: Breakdown of scatter plots from Supplementary Figures 4.3 and 4.9 in the same style as Supplementary Figure 4.19. (a) Supplementary Figure 4.15. The first row is DAPI prediction. The second row is CellMask prediction. (b) Supplementary Figure 4.16.

Chapter 5

Dermatology

5.1 Dermatologist-level Classification of Skin Cancer with Deep Neural Networks

Skin cancer - the most common human malignancy [1] [352] [401] - is primarily diagnosed visually, beginning with an initial clinical screening, followed potentially by dermoscopic analysis, a biopsy, and histopathological examination. Automated classification of skin lesions using images is a challenging task due to the fine-grained variability of skin lesion appearance. Deep convolutional neural networks (CNN) [246] [248] show great promise for general and highly variable tasks over many fine-grained object categories [361] [233] [189] [410] [411] [171]. Here we show classification of skin lesions using a single CNN, trained end-to-end directly from images using only their pixels and disease labels as inputs. We train a CNN on a dataset of 129,450 clinical images - two orders of magnitude larger than previous datasets [275] - consisting of 2,032 different diseases. We test its performance against 21 board-certified dermatologists on biopsy-proven clinical images with two critical use cases: binary classification of (1) malignant carcinomas versus benign seborrheic keratoses, and (2) malignant melanomas versus benign nevi. Case (1) represents the identification of the most common cancers, and case (2) represents identification of the deadliest skin cancer. The CNN achieves performance on par with all tested experts across both tasks, demonstrating, for the first time, an artificial intelligence with dermatologist-level skin cancer classification capability. It is projected that 6.3 billion smartphone subscriptions will exist by the year 2021 [77]. Outfitted with deep neural networks, mobile devices can extend the reach of dermatologists outside of the clinic, and enable low-cost universal access to vital diagnostic care.

There are 5.4 million new cases of skin cancer each year in the United States [352]. One in five Americans will be diagnosed with a cutaneous malignancy in their lifetime. While melanomas represent fewer than 5% of all skin cancers in the United States, they account for approximately 75% of all skin cancer-related deaths, and are responsible for over 10,000 deaths annually in the United States alone. Early detection is critical - the estimated 5-year survival rate of melanoma drops from 97% if detected in its earliest stages to 14% if detected in its latest stages. The key contribution of this work is a computational method which may allow medical practitioners and patients to proactively track skin lesions and detect cancer earlier. By creating a novel disease taxonomy and partitioning algorithm which map individual diseases into training classes, we build a deep learning system for automated dermatology.

Prior work in dermatological computer-aided classification [275] [356] [60] has lacked the generalization capability of medical practitioners due to insufficient data and a focus on standardized tasks such as dermoscopy [219] [91] [156] and histology image classification [37] [279] [90] [374]. Dermoscopy images are acquired via a specialized instrument and histology images are acquired via invasive biopsy and microscopy; both modalities yield highly standardized images. Photographic images (e.g. smartphone images) exhibit variability in zoom, angle, lighting, etc, making classification significantly more challenging [337] [22]. We overcome this challenge using a data-driven approach - 1.41 million pre-training and training images make it

robust to photographic variability. Many former techniques require extensive preprocessing, lesion segmentation, and extraction of domain-specific visual features prior to classification. In contrast, our system requires no hand-crafted features - it is trained end-to-end directly from image labels and raw pixels, with a single network for both photographic and dermoscopic images. The existing body of work uses small datasets of typically less than a thousand skin lesion images [156] [219] [37] that, as a result, do not generalize well to new images. We demonstrate generalizable classification with a new dermatologist-labeled dataset of 129,450 clinical images, including 3,374 dermoscopy images.

Deep learning algorithms, powered by advances in computation and extremely large datasets [102], have recently been shown to exceed human performance at visual AI tasks such as Atari game playing [292], strategic board games like Go [387], and object recognition6. In this paper we outline the development of a CNN that matches the performance of dermatologists at three key diagnostic tasks: melanoma classification, melanoma classification using dermoscopy, and carcinoma classification. We restrict comparison to image-based classification.

We utilize a GoogleNet Inception-v3 CNN architecture [410] pretrained on 1.28 million images (1,000 object categories) from the 2014 ImageNet Large Scale Visual Recognition Challenge6, and train it on our dataset using transfer learning [318]. Figure 5.1 demonstrates the working system.

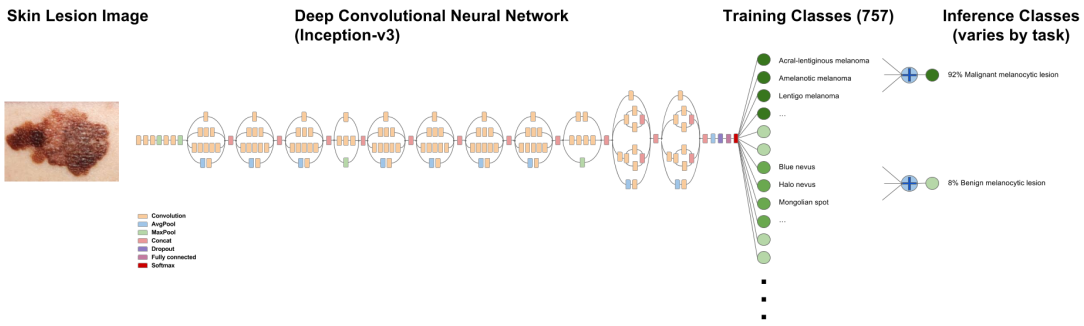


Figure 5.1: **Deep convolutional neural network layout.** Our classification technique is a deep convolutional neural network (CNN). Data flow is from left to right: an image of a skin lesion (e.g. melanoma) is sequentially warped into a probability distribution over clinical classes of skin disease using Googles Inception-v3 CNN architecture pretrained on the ImageNet dataset (1.28 million images over 1000 generic object classes) and fine-tuned on our own dataset of 129,450 skin lesions comprised of 2,032 different diseases. 757 training classes are defined using a novel taxonomy of skin disease and a partitioning algorithm that maps diseases into training classes (e.g. acrolentiginous melanoma, amelanotic melanoma, lentigo melanoma). Inference classes are more general and are composed of one or more training classes (e.g. malignant melanocytic lesions - the class of melanomas). The probability of an inference class is calculated by summing the probabilities of the training classes according to taxonomy structure.
*** Inception-v3 CNN architecture reprinted from https://research.googleblog.com/2016/03/train-your-own-image-classifier-with.html*

The CNN is trained using 757 disease classes. Our dataset is composed of dermatologist-labeled images

organized in a novel tree-structured taxonomy of 2,032 diseases, where the individual diseases form the leaf nodes. The images come from 18 different clinician-curated, open-access online repositories, as well as clinical data from Stanford Hospital and Clinics. Figure 5.2(a) shows a subset of the full taxonomy, which has been organized clinically and visually by medical experts. We split our dataset into 127,463 training/validation images, and 1,942 biopsy-labeled test images.

To take advantage of fine-grained information contained in the taxonomy structure, we develop an algorithm (Table 5.6) to partition diseases into fine-grained training classes (e.g. amelanotic melanoma and acrolentiginous melanoma). During inference, the CNN outputs a probability distribution over these fine classes. To recover the probabilities for coarser-level classes of interest (e.g. melanoma), we sum the probabilities of their descendants. See Methods and Extended Data Figure 5.1 for more details.

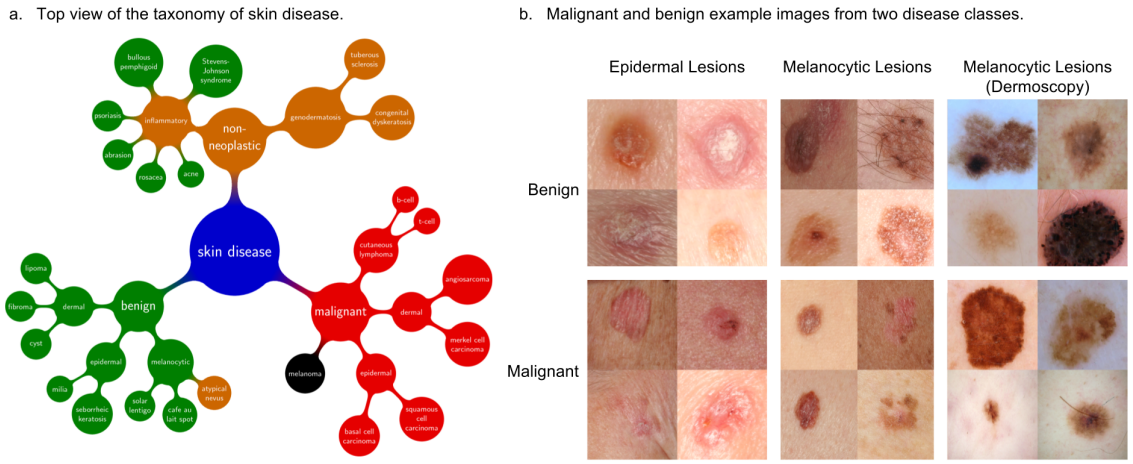


Figure 5.2: **A schematic illustration of the taxonomy and example test set images. a.** A subset of the top of the tree-structured taxonomy. The full taxonomy contains 2,032 diseases and is organized based on visual and clinical similarity of diseases. Red indicates malignant, green indicates benign, and orange indicates conditions that can be either. Black is melanoma. The first two levels of the taxonomy are used in validation. Testing is restricted to the tasks of b. **b.** Example test set images highlight the difficulty of malignant vs benign discernment for the three medically-critical classification tasks we consider: epidermal lesions, melanocytic lesions, and melanocytic lesions visualized with a dermoscope.

We validate the algorithms effectiveness in two ways, using nine-fold cross validation. First, we validate with a three-class disease partition - the level-1 nodes of the taxonomy - representing benign lesions, malignant lesions, and non-neoplastic lesions. Here the CNN achieves $72.1 \pm 0.9\%$ overall accuracy (the average of individual inference class accuracies) and two dermatologists attain 65.56% and 66.0% accuracy on a subset of the validation set. Second, we validate with a nine-class disease partition - the level-2 nodes - such that the diseases of each class have similar medical treatment plans. The CNN achieves $55.4 \pm 1.7\%$ overall accuracy whereas the same two dermatologists attain 53.3% and 55.0% accuracy. A CNN trained on a finer disease partition performs better than one trained directly on three or nine classes (see Table 2), demonstrating the

effectiveness of our partitioning algorithm. Since validation set images are labeled by dermatologists but not necessarily biopsy-proven, this metric is inconclusive, and instead shows that the CNN is learning relevant information.

For conclusiveness, we test our algorithm and dermatologists using strictly biopsy-proven images on medically significant use cases: distinguishing malignant versus benign (1) epidermal lesions (malignant carcinoma vs benign seborrheic keratosis) and (2) melanocytic lesions (malignant melanoma vs benign nevi). For (2), we display two trials, one using standard images and the other using dermoscopy images, which reflect two steps that a dermatologist might pursue to obtain a clinical impression. The same CNN is used across all three tasks. Figure 5.2(b) shows a few example images, demonstrating the difficulty in distinguishing between malignant and benign lesions, which share many visual features. Our comparison metrics are sensitivity and specificity (SS):

$$\text{sensitivity} = \frac{TP}{P}$$
$$\text{specificity} = \frac{TN}{N}$$

where TP (true positives) is the number of correctly predicted malignant lesions, P is the number of malignant lesions shown, TN (true negatives) is the number of correctly predicted benign lesions, and N is the number of benign lesions shown. When a test set is fed through the CNN, it outputs a probability, p, of malignancy, per image. We can compute the sensitivity and specificity of these probabilities by choosing a threshold probability t such that the prediction y for each image is given by $y = p > t$. Varying t in the interval $[0,1]$ generates a curve of sensitivities and specificities that the CNN can achieve.

Figure 5.4(a) shows a direct performance comparison between the CNN and over 21 board-certified dermatologists on epidermal and melanocytic lesion classification. For each image the dermatologists are asked whether to biopsy/treat the lesion or reassure the patient. Each red point on the plots represents the SS of a single dermatologist. The CNN outperforms any dermatologist whose SS point falls below the CNNs blue curve - most do. The green points represents the average dermatologist (average SS of all red points), with error bars denoting one standard deviation. The area-under-the-curve (AUC) for each case is over 91%. The data for this comparison (135 epidermal, 130 melanocytic, and 111 melanocytic-dermoscopy images) are sampled from the full test sets. Plotted in Figure 5.4(b) are the SS curves for our entire test set of biopsy-labeled images comprised of 707 epidermal, 225 melanocytic, and 1,010 melanocytic-dermoscopy images. From Figure 5.4(a) to Figure 5.4(b) we observe negligible changes in AUC (< 0.03), validating the reliability of our results on a larger dataset. In a separate analysis with similar results (see below) dermatologists are asked if they believe a lesion is malignant or benign.

Using t-SNE [262], we examine in Figure 5.5 the internal features learned by the CNN. Each point represents a skin lesion image projected from the 2048-dimensional output of the CNNs last hidden layer into two dimensions. We see clusters of points of the same clinical classes - insets show images of different diseases.

a.

Classifier	Three-way accuracy
Dermatologist 1	65.6%
Dermatologist 2	66.0%
CNN	69.4 ± 0.8%
CNN - PA	**72.1 ± 0.9%**

b.

Classifier	Nine-way accuracy
Dermatologist 1	53.3%
Dermatologist 2	55.0%
CNN	48.9 ± 1.9%
CNN - PA	**55.4 ± 1.7%**

c. **Disease classes: three-way classification**

0. Benign single lesions
1. Malignant single lesions
2. Non-neoplastic lesions

d. **Disease classes: nine-way classification**

0. Cutaneous lymphoma and lymphoid infiltrates
1. Benign dermal tumors, cysts, sinuses
2. Malignant dermal tumor
3. Benign epidermal tumors, hamartomas, milia, and growths
4. Malignant and premalignant epidermal tumors
5. Genodermatoses and supernumerary growths
6. Inflammatory conditions
7. Benign melanocytic lesions
8. Malignant Melanoma

Figure 5.3: **General Validation Results** Here we show ninefold cross-validation classification accuracy with 127,463 images organized in two different strategies. In each fold, a different ninth of the dataset is used for validation, and the rest is used for training. Reported values are the mean and standard deviation of the validation accuracy across all n = 9 folds. These images are labelled by dermatologists, not necessarily through biopsy; meaning that this metric is not as rigorous as one with biopsy-proven images. Thus we only compare to two dermatologists as a means to validate that the algorithm is learning relevant information. **a**, Three-way classification accuracy comparison between algorithms and dermatologists. The dermatologists are tested on 180random images from the validation set 60 per class. The three classes used are first-level nodes of our taxonomy. A CNN trained directly on these three classes also achieves inferior performance to one trained with our partitioning algorithm (PA). **b**, Nine-way classification accuracy comparison between algorithms and dermatologists. The dermatologists are tested on 180 random images from the validation set 20 per class. The nine classes used are the second-level nodes of our taxonomy. A CNN trained directly on these nine classes achieves inferior performance to one trained with our partitioning algorithm. **c**, Disease classes used for the three-way classification represent highly general disease classes. **d**, Disease classes used for nine-way classification represent groups of diseases that have similar aetiologies.

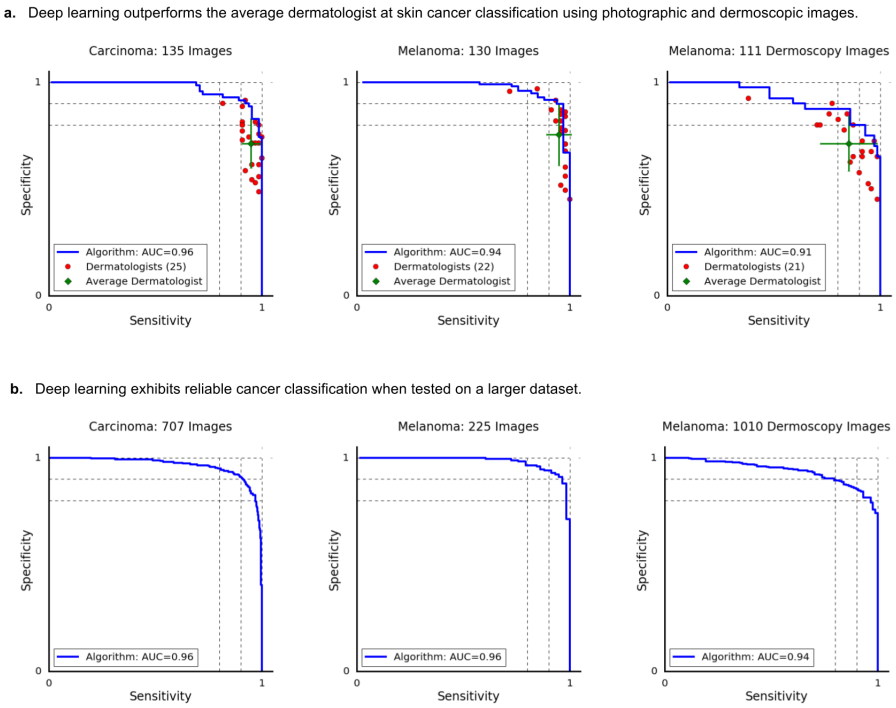


Figure 5.4: **Skin cancer classification performance of the CNN and dermatologists.** a, The deep learning CNN outperforms the average of the dermatologists at skin cancer classification using photographic and dermoscopic images. Our CNN is tested against at least 21dermatologists at keratinocyte carcinoma and melanoma recognition. For each test, previously unseen, biopsy-proven images of lesions are displayed, and dermatologists are asked if they would: biopsy/treat the lesion or reassure the patient. Sensitivity, the true positive rate, and specificity, the true negative rate, measure performance. A dermatologist outputs a single prediction per image and is thus represented by a single red point. The green points are the average of the dermatologists for each task, with error bars denoting one standard deviation (calculated from n = 25, 22 and 21 tested dermatologists for keratinocyte carcinoma, melanoma and melanoma under dermoscopy, respectively). The CNN outputs a malignancy probability P per image. We fix a threshold probability t such that the prediction $\check{y}$ for any image is $\check{y} = P \geq t$, and the blue curve is drawn by sweeping t in the interval 01. The AUC is the CNNs measure of performance, with a maximum value of 1. The CNN achieves superior performance to a dermatologist if the sensitivityspecificity point of the dermatologist lies below the blue curve, which most do. Epidermal test: 65 keratinocyte carcinomas and 70 benign seborrheic keratoses. Melanocytic test: 33 malignant melanomas and 97 benign nevi. A second melanocytic test using dermoscopic images is displayed for comparison: 71 malignant and 40 benign. The slight performance decrease reflects differences in the difficulty of the images tested rather than the diagnostic accuracies of visual versus dermoscopic examination. b, The deep learning CNN exhibits reliable cancer classification when tested on a larger dataset. We tested the CNN on more images to demonstrate robust and reliable cancer classification. The CNNs curves are smoother owing to the larger test set.

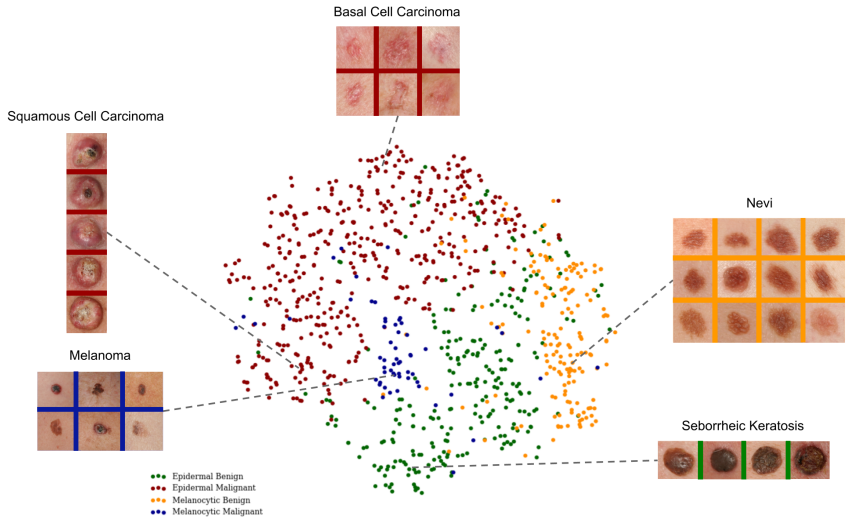


Figure 5.5: **t-SNE visualization of the last hidden layer representations in the CNN for four disease classes.** Here we show the CNNs internal representation of four important disease classes by applying t-SNE, a method for visualizing high-dimensional data, to the last hidden layer representation in the CNN of the biopsy-proven photographic test sets (932 images). Coloured point clouds represent the different disease categories, showing how the algorithm clusters the diseases. Insets show images corresponding to various points. Images reprinted with permission from the Edinburgh Dermofit Library (https://licensing.eri.ed.ac.uk/i/

Basal and squamous cell carcinomas are split across the malignant epidermal point cloud. Melanomas lie in the center, in contrast to nevi, on the right. Similarly, seborrheic keratoses lie across from their malignant counterparts.

Here we demonstrate the effectiveness of deep learning in dermatology - a technique which we apply to both general skin conditions and specific cancers. Using a single convolutional neural network trained at general skin lesion classification, we match the performance of over 21 tested dermatologists across three critical diagnostic tasks: carcinoma classification, melanoma classification, and melanoma classification under dermoscopy. This fast, scalable method will be deployable on mobile devices and holds the potential for significant clinical impact, including broadening the scope of primary care practice and augmenting clinical decision-making for dermatology specialists. Further research is necessary to evaluate performance in a real-world, clinical setting to validate this technique across the full distribution and spectrum of lesions encountered in typical practice. Acknowledging that a dermatologists clinical impression and diagnosis is based on contextual factors beyond visual and dermoscopic inspection of a lesion in isolation, the ability to classify skin lesion images with the accuracy of a board-certified dermatologist has the potential to dramatically expand access to vital medical care. This method is primarily constrained by data and can classify many visual conditions if sufficient training examples exist. Deep learning is agnostic to the type of image data used and could be adapted to other specialties, including ophthalmology, otolaryngology, radiology, and pathology.

5.1.1 Datasets

Our dataset comes from a combination of open-access dermatology repositories, the ISIC Dermoscopic Archive, the Edinburgh Dermofit Library [374], and data from the Stanford Hospital. The images from the online open-access dermatology repositories are annotated by dermatologists, not necessarily through biopsy. The ISIC Archive data used is composed strictly of melanocytic lesions that are biopsy-proven and annotated as malignant or benign. The Edinburgh Dermofit Library and data from the Stanford Hospital are biopsy-proven and annotated by individual disease names (i.e. actinic keratosis). In our test sets, melanocytic lesions include malignant melanomas - the deadliest skin cancer - and benign nevi. Epidermal lesions include malignant basal and squamous cell carcinomas, intraepithelial carcinomas, pre-malignant actinic keratosis, and benign seborrheic keratosis.

5.1.2 Taxonomy

Our taxonomy represents 2,032 individual diseases arranged in a tree structure with its three root nodes representing general disease classes: (1) benign lesions, (2) malignant lesions, and (3) non-neoplastic lesions (Figure 5.2(b)). It was derived by dermatologists using a bottom-up procedure: individual diseases, initialized as leaf nodes, were merged based on clinical and visual similarity, until the entire structure was connected. This aspect of the taxonomy makes it useful in generating training classes that are both well-suited for machine learning classifiers and medically relevant. The root nodes are used in the first validation strategy and represent the most general partition. The children of the root nodes (i.e. malignant melanocytic lesions) are used in the second validation strategy, and represent disease classes that have similar clinical treatment plans.

5.1.3 Data Preparation

Blurry images and far-away images were removed from the test and validation sets, but still used in training. Our dataset contains sets of images corresponding to the same lesion but from multiple viewpoints, or multiple images of similar lesions on the same person. While this is useful training data, extensive care was taken to ensure that these sets were not split between the training and validation sets. Using image EXIF metadata, repository specific information, and nearest neighbor image retrieval with CNN features, we created an undirected graph connecting any pair of images that were determined to be similar. Connected components of this graph were not allowed to straddle the train/validation split and were randomly assigned to either train or validation. The test sets all came from independent, high-quality repositories of biopsy-proven images - the Stanford Hospital, the University of Edinburgh Dermofit Image Library, and the ISIC Dermoscopic Archive. No overlap (i.e. same lesion multiple viewpoints) exists between the test sets and the training/validation data.

5.1.4 Sample Selection

The epidermal, melanocytic, and melanocytic-dermoscopic tests of Figure 5.4(a) used 135 (65 malignant, 70 benign), 130 (33 malignant, 97 benign), and 111 (71 malignant, 40 benign) images, respectively. Their

counterparts of Figure 5.4b used 707 (450 malignant, 257 benign), 225 (58 malignant, 167 benign), and 1010 (88 malignant, 922 benign) images, respectively. The number of images used for Figure 5.4(b) was based on the availability of biopsy-labeled data (i.e. malignant melanocytic lesions are exceedingly rare compared to benign melanocytic lesions). These numbers are statistically justified by the standards of the ILSVRC computer vision challenge [361], which has 50-100 images per class for validation and test sets. For (a), 140 images were randomly selected from each set of (b), and a non-tested dermatologist (blinded to diagnosis) removed any images of insufficient resolution (while the network accepts 299 x 299 image inputs, humans require larger images for clarity).

5.1.5 Disease Partitioning Algorithm

The algorithm that partitions the individual diseases into training classes is outlined more formally in Figure 5.6. It is a recursive algorithm, designed to leverage the taxonomy to generate training classes whose individual diseases are clinically and visually similar. The algorithm forces the average generated training class size to be slightly less than its only hyperparameter, $maxClassSize$. Together these components strike a balance between (1) generating training classes that are overly-fine grained and dont have sufficient data to be learned properly, (2) generating training classes that are too coarse, too data abundant, and bias the algorithm towards them. With $maxClassSize = 1000$ this algorithm yields a disease partition of 757 classes. All training classes are descendants of inference classes.

5.1.6 Training Algorithm

We use Googles Inception-v3 CNN architecture pre-trained to 93.33% top-5 accuracy on the 1000 object classes (1.28M images) of the 2014 ImageNet Challenge following Szegedy et al. [411]. We then remove the final classification layer from the network and retrain it with our dataset, fine-tuning the parameters across all layers. During training we resize each image to 299 x 299 pixels in order to make it compatible with the original dimensions of the Inception-v3 network architecture and leverage the natural-image features learned by the ImageNet pretrained network. This procedure, known as transfer learning, is optimal given the amount of data available.

Our CNN is trained using backpropagation. All layers of the network are fine-tuned using the same global learning rate of 0.001 and a decay factor of 16 every 30 epochs. We use RMSProp with decay of 0.9, momentum of 0.9, and epsilon of 0.1. We use Googles TensorFlow30 deep learning framework to train, validate, and test our network. During training, images are augmented by a factor of 720. Each image is rotated randomly between 0 and 359 degrees. The largest upright inscribed rectangle is then cropped from the image, and is flipped vertically with a probability of 1/2.

```
Disease Partitioning Algorithm
 1: Inputs
 2:       taxonomy (tree): the disease taxonomy
 3:       maxClassSize (int): maximum data points in a class
 4: Output
 5:       partition (list of sets): partition of the diseases into classes
 6:
 7: procedure DESCENDANTS(node)
 8:     return {node} ∪ {DESCENDANTS(child) for child in node.children}
 9:
10: procedure NUMIMAGES(nodes)
11:     return SUM(LENGTH(node.images) for node in nodes)
12:
13: procedure PARTITIONDISEASES(node)
14:     class ← DESCENDANTS(node)
15:     if NUMIMAGES(class) < maxClassSize then
16:         append class to partition
17:     else
18:         for child in node.children do
19:             PARTITIONDISEASES(child)
20:
21: partition ← [ ]
22: PARTITIONDISEASES(taxonomy.root)
23: return partition
```

Figure 5.6: **Disease-partitioning algorithm.** This algorithm uses the taxonomy to partition the diseases into fine-grained training classes. We find that training on these finer classes improves the classification accuracy of coarser inference classes. The algorithm begins with the top node and recursively descends the taxonomy (line 19), turning nodes into training classes if the amount of data contained in them (with the convention that nodes contain their children) does not exceed a specified threshold (line 15). During partitioning, the recursive property maintains the taxonomy structure, and consequently, the clinical similarity between different diseases grouped into the same training class. The data restriction (and the fact that training data are fairly evenly distributed amongst the leaf nodes) forces the average class size to be slightly less than *maxClassSize*. Together these components generate training classes that leverage the fine-grained information contained in the taxonomy structure while striking a balance between generating classes that are overly fine-grained and do not have sufficient data to be learned properly, and classes that are too coarse, too data abundant and that prevent the algorithm from properly learning less data-abundant classes. With $maxClassSize = 1,000$ this algorithm yields 757 training classes.

5.1.7 Inference Algorithm

We follow the convention that each node contains its children. Each training class is represented by a node in the taxonomy, and subsequently, all descendants. Each inference class is a node which has as its descendants some set of training nodes. An illustrative example is shown in Figure 5.7, with red nodes as inference classes and green nodes as training classes.

Given an input image, the CNN outputs a probability distribution over the training nodes. Probabilities over the taxonomy follow:

$$P(u) = \sum_{v \in C(u)} P(v) \tag{5.1}$$

Where u is any node, P(u) is the probability of u, and C(u) are the child nodes of u. Thus to recover the probability of any inference node we simply sum the probabilities of its descendant training nodes. Note that in the validation strategies all training classes are summed into inference classes. However in the binary

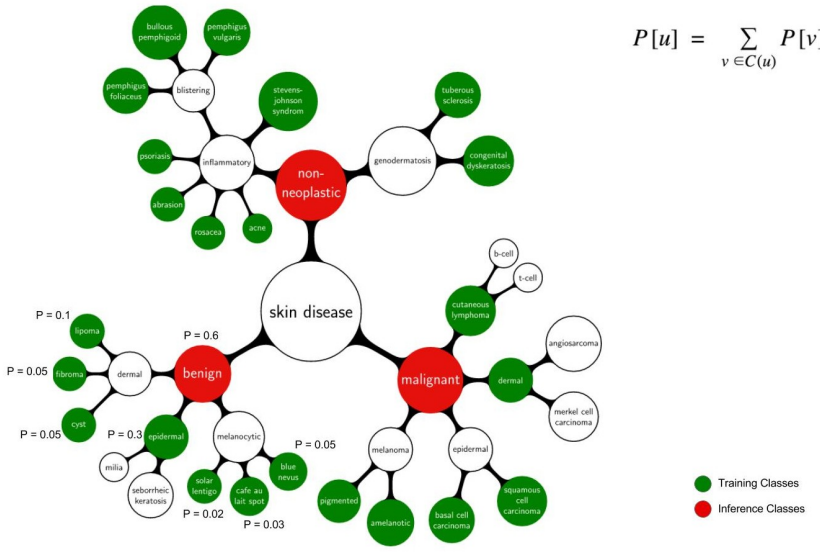


Figure 5.7: **Procedure for calculating inference class probabilities from training class probabilities.** Illustrative example of the inference procedure using a subset of the taxonomy and mock training/inference classes. Inference classes (for example, malignant and benign lesions) correspond to the red nodes in the tree. Training classes (for example, amelanotic melanoma, blue nevus), which were determined using the partitioning algorithm with $maxClassSize = 1{,}000$, correspond to the green nodes in the tree. White nodes represent either nodes that are contained in an ancestor nodes training class or nodes that are too large to be individual training classes. The equation represents the relationship between the probability of a parent node, u, and its children, $C(u)$; the sum of the child probabilities equals the probability of the parent. The CNN outputs a distribution over the training nodes. To recover the probability of any inference node it therefore suffices to sum the probabilities of the training nodes that are its descendants. A numerical example is shown for the benign inference class: $P_{benign} = 0.6 = 0.1 + 0.05 + 0.05 + 0.3 + 0.02 + 0.03 + 0.05$.

classification cases, the images in question are known to be either melanocytic or epidermal and so we utilize only the training classes which are descendants of either melanocytic or epidermal.

5.1.8 Confusion Matrices

Figure 5.8 shows the confusion matrix of our method over the nine classes of the second validation strategy (Figure 5.3(d)) in comparison to the two tested dermatologists.

This demonstrates the misclassification similarity between the CNN and human experts. Element (i, j) of each confusion matrix represents the empirical probability of predicting class j given that the ground truth was class i. Classes 7 and 8 - benign and malignant melanocytic lesions - are often confused for each other. Many images are mistaken as class 6, the inflammatory class, due to the high variability of diseases in this category. Note how easily malignant dermal tumors are confused for other classes, by both the CNN and dermatologists. They are essentially nodules under the skin that are challenging to visually diagnose.

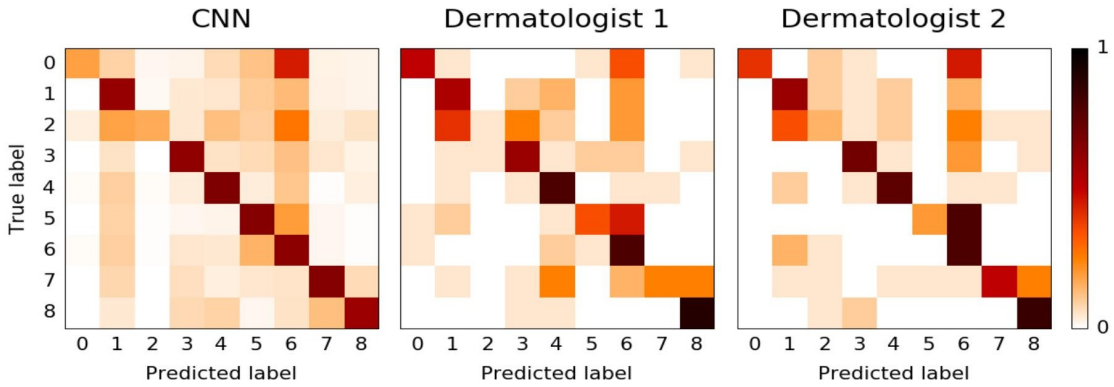


Figure 5.8: **Confusion matrix comparison between CNN and dermatologists.** Confusion matrices for the CNN and both dermatologists for the nine-way classification task of the second validation strategy reveal similarities in misclassification between human experts and the CNN. Element (i, j) of each confusion matrix represents the empirical probability of predicting class j given that the ground truth was class i, with i and j referencing classes from Extended Data Table 2d. Note that both the CNN and the dermatologists noticeably confuse benign and malignant melanocytic lesionsclasses 7 and 8with each other, with dermatologists erring on the side of predicting malignant. The distribution across column 6inflammatory conditionsis pronounced in all three plots, demonstrating that many lesions are easily confused with this class. The distribution across row 2 in all three plots shows the difficulty of classifying malignant dermal tumours, which appear as little more than cutaneous nodules under the skin. The dermatologist matrices are each computed using the 180images from the nine-way validation set. The CNN matrix is computed using a random sample of 684images (equally distributed across the nine classes) from the validation set.

5.1.9 Saliency Maps

To visualize the pixels that a network is fixating on for its prediction, we generate saliency maps, shown in Figure 5.9, for example images of the nine classes of Table 5.3(d). Backpropagation is an application of the chain rule of calculus to compute loss gradients for all weights in the network. The loss gradient can also be backpropagated to the input data layer. By taking the L_1 norm of this input layer loss gradient across the RGB channels, the resulting heat map intuitively represents the importance of each pixel for diagnosis. As can be seen, the network fixates most of its attention on the lesions themselves and ignores background and healthy skin.

5.1.10 Sensitivity-Specificity Curves with different question

In the main text we compare our CNNs SS to that of over 21 dermatologists on the three diagnostic tasks of Figure 5.4. In that analysis each dermatologist was asked if they would (a) biopsy/treat the lesion, or (b) reassure the patient. This choice of question reflects the actual in-clinic task that dermatologists must perform - deciding whether or not to continue medically analyzing a lesion. A similar question to ask a dermatologist, though less clinically relevant, is if they believe a lesion is (a) malignant, or (b) benign. The results of this

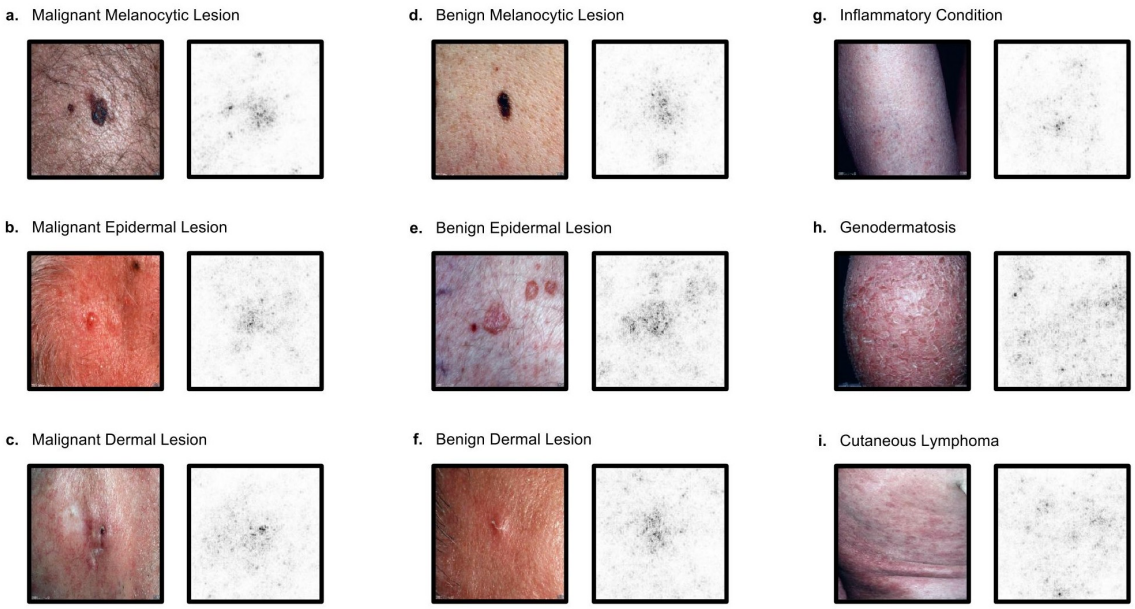


Figure 5.9: **Saliency maps for nine example images from the second validation strategy.** Saliency maps for example images from each of the nine clinical disease classes of the second validation strategy reveal the pixels that most influence a CNNs prediction. Saliency maps show the pixel gradients with respect to the CNNs loss function. Darker pixels represent those with more influence. We see clear correlation between the lesions themselves and the saliency maps. Conditions with single lesions - a, b, c, d, e, f - tend to exhibit tight saliency maps centered around the lesions themselves. Conditions with spreading lesions - g, h, i - exhibit saliency maps that similarly occupy multiple points of interest in the images. (a) malignant melanocytic lesion (b) malignant epidermal lesion (c) malignant dermal lesion (d) benign melanocytic lesion (e) benign epidermal lesion (f) benign dermal lesion (g) inflammatory condition (h) genodermatosis (i) cutaneous lymphoma.

analysis are shown in Figure 5.10. As in the main Figure 5.4, the CNN is on par with the performance of the dermatologists and outperforms the average. In the epidermal lesions test, the CNN is just above one standard deviation above the average dermatologist, and in both melanocytic lesion tests the CNN is just below one standard deviation above the average dermatologist.

Use of human subjects

All human subjects were board-certified dermatologists that took our tests under informed consent. This study was approved by the Stanford Institutional Review Board, under trial registration number 36050.

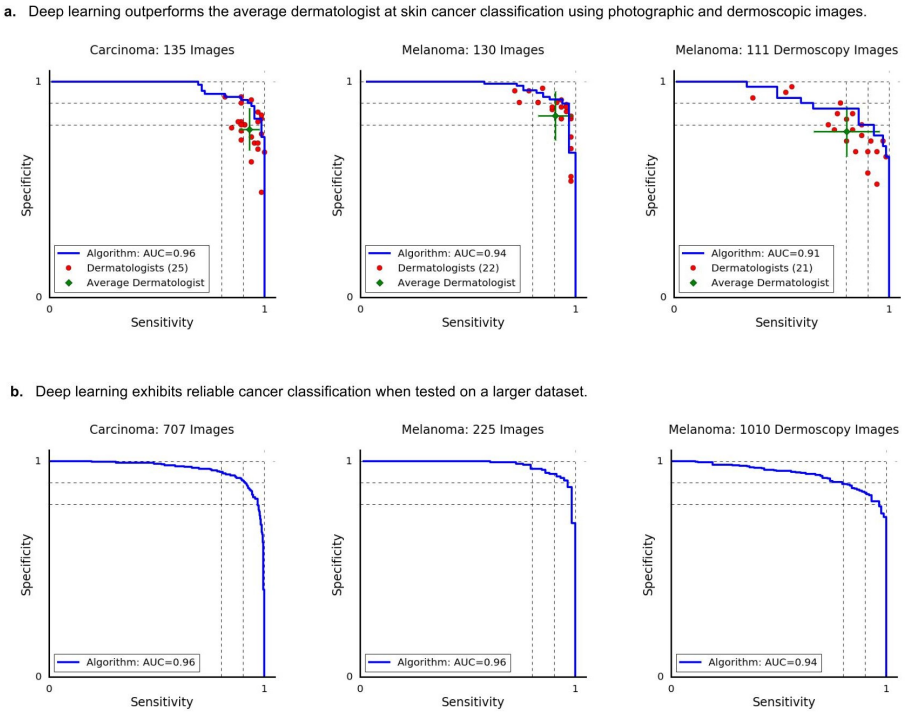


Figure 5.10: **Extension of Figure 3 with a different dermatological question.** a. Identical plots and results to Figure 3(a), except that dermatologists are asked if a lesion appears (a) malignant, or (b) benign. This is a somewhat unnatural question to ask - in-clinic, the only actionable decision is whether or not to biopsy/treat a lesion. The blue curves for the CNN are identical to Figure 3. b. Figure 3(b) reprinted for visual comparison to a.

5.1.11 Data Availability Statement

The medical test sets that support the findings of this study are available from the ISIC Archive (https://isic-archive.com/) and the Edinburgh Dermofit Library (https://licensing.eri.ed.ac.uk/i/software/dermofit-image-library.html). Restrictions apply to the availability of the medical training/validation data, which were used under permission for the current study, and so are not publicly available. Some data may be available from the authors upon reasonable request and with permission of the Stanford Hospital.

5.2 Skin Cancer Detection & Tracking with Deep Learning and Data Synthesis

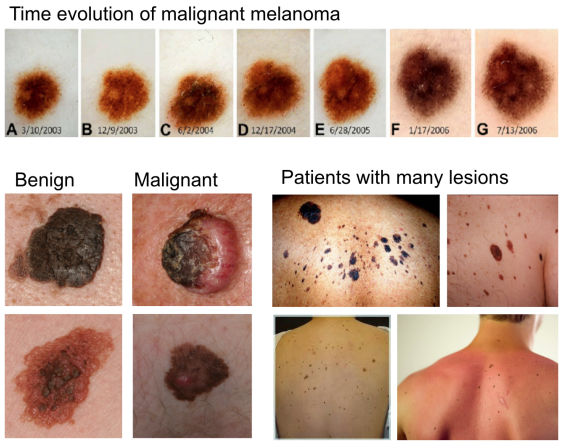


Figure 5.11: Key factors for skin cancer care include early detection and tracking over time. Top Row: superficial spreading melanoma, evolving in time [367]. Bottom left: comparison between malignant and benign lesions shows the difficulty in early detection. Bottom right: examples of patients afflicted with many lesions.

Dermatology is a medical field which stands to be heavily augmented by the use of artificial intelligence techniques. Diseases are visually screened for, and many disease diagnoses are performed strictly with an in-clinic visual examination. Discerning between skin lesions is a difficult task - the difference between skin cancer (melanoma, carcinoma) and benign lesions (nevi, seborrheic keratosis) is minute, and the differences are in slight details between them (Figure 5.11). With 5.4 million cases of skin cancer diagnosed each year in the United States alone, afflicting 3.3 million people, the need for quick and effective clinical screenings is rising [351, 400]. Patients with skin cancer tend to be afflicted with many moles, and so one of the challenges in skin cancer screenings is identifying them amongst a myriad of benign lesions. A key element of these diagnoses is based on inspecting temporal changes in lesions - a fast changing lesion is more likely to be malignant. As such, patients and providers need tools to support this at scale.

Recent advances in detection and tracking using CNNs [142, 141, 344, 196, 206, 478] has the potential to augment healthcare providers by (1) detecting points of malignancy, and (2) finding corresponding lesions across images, allowing them to be tracked temporally. However, the primary challenge in using traditional detection techniques is working in a low-data regime without the availability of high volumes of annotated and labeled data - the largest existing open-source skin cancer dataset of photographic images is the Edinburgh Dermofit dataset, containing 1,300 biopsied images. It is worth noting that a larger dataset, the ISIC Archive, exists, but contains dermoscopic images which are not relevant in large-scale detection and tracking. Dermoscopic images are taken with a specialized instrument known as a dermoscope which yields

highly standardized images of single lesions. The size of a dermoscope is on the order of a lesion and thus is not suitable for distanced images which capture many lesions and potentially entire areas of a person's bare skin.

To overcome the challenge of working in this low-data regime we develop a domain-specific data synthesis technique which stitches small single-lesion images onto large body images. Both the body images and the skin lesions images are heavily augmented with various techniques, and the lesions are blended onto the bodies using Poisson image editing.

For large-scale lesion detection, we use this synthetic data to train a fully-convolutional CNN on the task of pixel-wise classification between three classes: background, benign lesion, malignant lesion. For image-to-image tracking, a network is trained (using image pairs containing altered lesion and pose positions) to output pixel-wise positional shift.

Our method demonstrates a working end-to-end CNN system capable of tackling two critical diagnostic tasks with superior performance to algorithmic baseline techniques. It is trained with very little original data and thus the techniques demonstrated here can be easily transferred to other data-limited domains. As an additional baseline we include a simple comparison between our model and two humans trained on the 1,300 images of the Edinburgh dataset.

Outfitted with CNNs, mobile devices could be used to take full-body images and detect and track lesions across them. By leveraging the 6.3 billion smart-phone subscriptions that are projected to exist by 2021 [76], healthcare could be extended outside the clinic and reach a much broader demographic.

5.2.1 Related Work

Detection

Prior work at the intersection of artificial intelligence and dermatology has focused largely on standardized tasks such as dermoscopy [92] and histology image classification and detection [38, 89]. Dermoscopy is a medical imaging modality involving a small device that sits on the skin and provides special illumination, and is used by physicians to more closely inspect suspicious lesions. Histology refers to tissue analysis of biopsied lesions using microscopy, and is considered the final step in a diagnosis of malignancy.

In contrast, our system is designed for photographic clinical supervision - it is intended as a precursor to these other screening techniques and could potentially aid a physician at spotting a suspicious lesion amongst benign ones. Prior techniques for photographic image analysis have also been constrained by data abundance [338] - a challenge we overcome using data synthesis and augmentation.

State-of-the-art techniques in image detection rely heavily on large datasets with annotated bounding boxes that tend to occupy significant portions of the image, whereas in this problem setting, the objects of interest occupy very small portions. Pixel-wise predictions have proven to be effective in targeting small areas such as facial keypoints [386, 252], and have gained popularity in semantic segmentation [257]. We leverage a combination of these techniques with heavy data augmentation to train a pixel-wise classifier and

apply post-processing to convert it into a detector.

Tracking

Feature-matching techniques using SIFT [255] and CNN features [258] show great promise for pixel-wise correspondence, and can be sped up with spatial pyramid techniques [215]. We expand this idea and develop task-specific supervised learning to enhance feature matching.

Training CNNs using synthetic data has proven effective at rigid-body correspondence matching [478], viewpoint estimation [406], and optical flow estimation [109]. However, many problems, including skin lesion detection, involve highly deformable and variable bodies where this data augmentation is less useful. We combine insights from these works to generate data and train CNNs that can handle variations in deformable bodies as well as large-scale positional changes.

5.2.2 Data Synthesis

We create a domain-specific data augmentation technique for generating synthetic images from two low-data sources: high-quality lesion images (1,300 biopsy-proven cancers and moles), and body images (400 back, leg, and chest images) whose skin regions have been manually segmented. We first generate images for detection and then further augment them for tracking. These images are intended to mimic the real-world clinical case of patients exhibiting many lesions, some possibly malignant, with the need to track them over time.

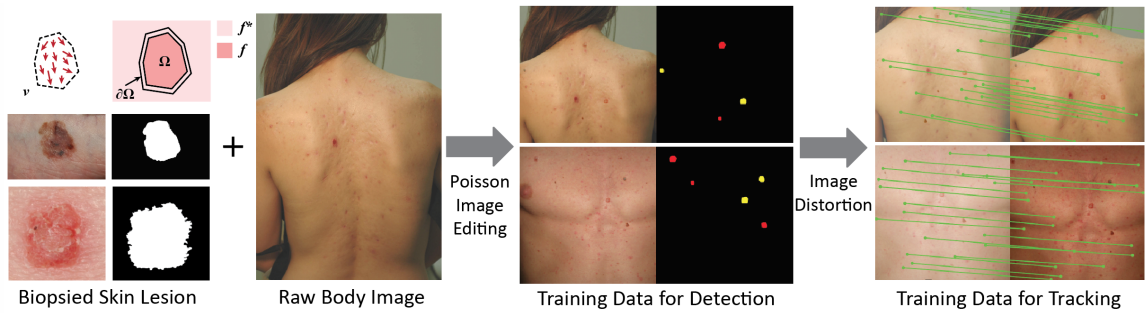


Figure 5.12: **Data Synthesis.** Skin lesions are blended with raw body images to generate detection and tracking data. (Left) Example biopsied skin lesion and raw body images. Top diagrams show the lesion segmentation mask and the gradient field along with semantic regions used to calculate blending locations. (Middle) Generated training images for detection and corresponding label masks. Red areas represent blended malignant lesions, yellow areas represent blended benign lesions. (Right) Generated training images for tracking, along with a few example pixel-wise correspondences.

Detection Data

Detection data is generated in two steps: (1) A blending position on the body image is chosen using local feature matching between a lesion image and the body image, (2) the lesion image is blended into the body image using Poisson image editing [326].

We select a random skin lesion image, re-size it such that its size is roughly on the order of the existing lesions of the body picture in question, randomly rotate it, and select a random position on the skin of a body image. The color histogram features of the surrounding area of the skin lesion (f^* in the top left corner of Figure 5.12) and the area of the same size as the skin lesion patch on the body image are used to determine if the lesion and position are a match. We use the earth mover's distance (EMD) as the comparison metric, and if it is lower than a predefined threshold, we blend the lesion at the chosen location, otherwise we randomly select another. Finally, we use Poisson Image Editing (see section below) to blend the lesion onto the chosen location. This process (and in particular, the use of color histogram features) minimizes the necessary change in color needed to seamlessly blend the lesion onto the image, qualitatively improving the final appearance of the blended images. Note that our choice is largely due to the model's simplicity and efficiency - many others exist. Example blending results along with the label mask of the attached lesions can be seen in Figure 5.12.

The basic idea behind Poisson image editing is to manipulate the intensity's gradient field between source and target instead of the absolute intensity. The advantage to doing this is that human perception is far more sensitive to differences in intensity than the values themselves. The segmentation masks provided with the Edinburgh dataset allow this technique to seamlessly blend lesion and body images together.

Let $\mathbf{v}$ be the gradient field of the source image (lesion), $\blacksquare$ the corresponding masked area in the target image (body), $\partial\blacksquare$ the boundary of the masked area, $\boldsymbol{f}^*$ the known function of the target image, and $\boldsymbol{f}$ the unknown function in area $\blacksquare$ that needs to be calculated, as shown in Figure 5.12. Then the problem can be formulated as optimizing:

$$\min_{\boldsymbol{f}} \iint_{\blacksquare} |\nabla \boldsymbol{f} - \mathbf{v}|^2 \text{ with } \boldsymbol{f}|_{\partial\blacksquare} = \boldsymbol{f}^*|_{\partial\blacksquare} \tag{5.2}$$

where $\nabla\cdot = [\frac{\partial\cdot}{\partial x}, \frac{\partial\cdot}{\partial y}]$ is the gradient operator, whose solution is the solution of a Poisson equation.

We generate pixel-wise labels using five classes for a subset of all pixels in the image. The attached lesions (the foreground) consist of four of these classes: malignant epidermal, benign epidermal, malignant melanocytic, and benign melanocytic lesions. The fifth class, the background, consists of healthy skin and non-skin image regions. Suppose that a given image has a total of N foreground pixels. Then we randomly sample $4N$ background pixels from the remainder of the image, manually excluding regions where the original body images contained obvious lesions. This balances the learning procedure - the algorithm strictly learns to detect biopsy-proven lesions without overcompensating towards background.

Tracking Data

For tracking, data is generated by further augmenting detection images - that is, given a detection image, we create a pair of images from it. The purpose here is to recreate temporal images (which will exhibit changes in lesion shape/size, as well as body changes) and to force the network to learn pixel-wise correspondence by focusing on the distortion in local texture information.

Skin lesions change as time passes. We simulate this by distorting the attached lesions on the body. We randomly adjust their shape, size, and brightness, and we randomly move them small distances. The coefficients of these transformation are all randomly set to increase diversity.

Brightness is distorted using gamma correction with randomly set coefficients. We apply elastic deformation as described in [389]. This is done by generating random displacement fields $\Delta x(x,y) = \text{rand}(-1,+1)$ and $\Delta y(x,y) = \text{rand}(-1,+1)$, generated with a uniform distribution, convolving it with a Gaussian of a manually decided standard deviation σ (15% of the smallest image dimension), and applying this displacement field to the image. We further distort the pose with a perspective transformation: four points are randomly sampled from each of the four corner boxes of the image (the height and width of each box being 1/6th of the height and width of the image), and the resultant quadrilateral is cropped and reshaped into the original height and width.

For each detection image, we generate one such distorted image, and then generate two tracking images by taking a random crop of the original image and an equal-sized crop of the distorted image, randomly translated between 0 and $1/10$th of the crop's dimensions. Sample images can be seen in Figure 5.12.

This procedure preserves pixel-wise correspondence between the two generated tracking images. This correspondence (a 2D pixel-wise displacement vector field) serves as the label for training our tracking network.

5.2.3 System Pipeline

Our system is composed of two parts: the first detects malignant and benign skin lesions, the second tracks them across images. The detection network is trained with the synthetic images (skin lesion + body images) described in the previous section, using pixel-wise labels. Once the detection network is trained to convergence, its weights are used to initialize the tracking network. This network is then trained on image-pairs formed from the detection data. It is worth noting that this particular choice of architecture is largely the result of iterating on potential architectures to determine the one with superior performance.

Detection System

The detection component is intended to highlight to a clinician the potentially malignant lesions on a given input image. Providers are often faced with body regions containing a multitude of lesions and discerning malignancy is a challenging task. Our system feed-forwards an input image through the CNN, outputs a pixel-wise heat-map over the five classes of interest, and then uses post-processing techniques to make the

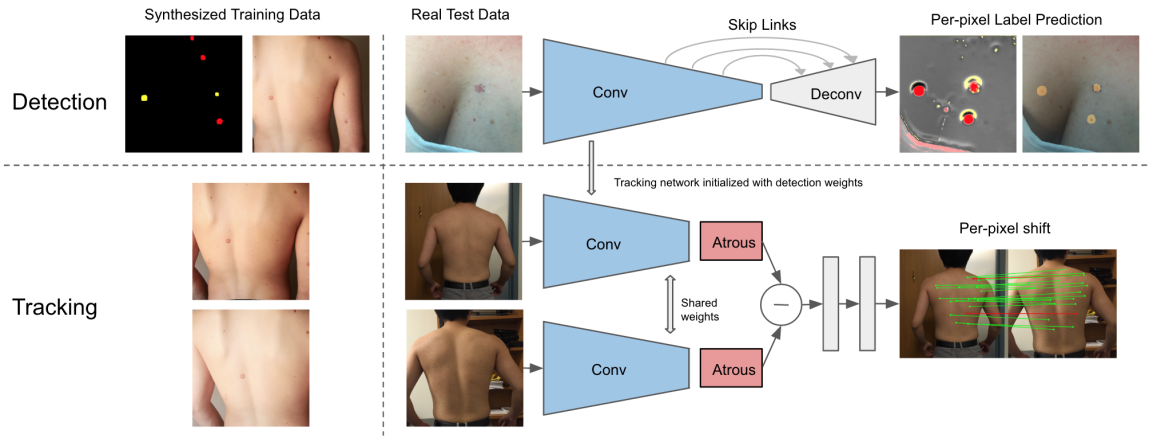


Figure 5.13: **Detection and Tracking System.** The network is trained on synthetic data and tested on real data. Top row shows the detection pipeline, bottom row shows tracking. The detection network is composed of a convolution section followed by a deconvolution section, with skip-link connections between non-adjacent layers. The network outputs per-pixel labels over two malignant, two benign, and one background class. In the top right we show the raw prediction heat map and the detection result after post processing. The tracking network takes the convolutional component of the detection network, and splits it up into a smaller convolutional part, and an atrous convolutional part. The two tracking images are each fed through the network and merged by a subtraction before the per-pixel shift prediction.

heat-map more human-interpretable.

The network structure, shown in Figure 5.13, is composed of a convolutional component and a deconvolutional component, a proven technique for pixel-wise prediction tasks [257, 309]. The design of the convolutional component has been adapted from Zimmerman's VGG16 network [390], which has been shown to be effective across several vision tasks [360, 247] We utilize the architecture of all layers up until `conv5_3` in VGG16, resulting in the spatial extent of the final feature map being 16 times smaller than that of the input image. In the deconvolution part of the network, we connect three groups of deconv-conv pairs to yield the final output. Each deconv-conv pair is composed of a deconvolutional layer which upsamples the feature map to twice its size, followed by a 3×3 VALID convolution (i.e. with padding) with a stride of 1, that does not further change the feature map size.

Additionally, we include skip-link connections as described in [355]. These component networks have proven to be effective at pixel-wise predictions and biomedical segmentation. The structure of our network can be viewed in Figure 5.13.

For an $M \times N \times 3$ input image, the network is trained to output a $M/2 \times N/2 \times 5$ heatmap (downscaled 16 times by the convolutional part and upscaled 8 times by the deconvolutional part), where each pixel $\mathbf{x}_i$ is a 5-dimensional vector representing the probability distribution over the 5 classes. Let i index the number of image pixels n, that were assigned labels, let the position and label pairs be $\{(\mathbf{x}_i, t_i)_{i=1}^{n}\}$, and let the pixel-wise probability distributions be written as as $p^{\mathbf{x}_i} = (p_0^{\mathbf{x}_i}, p_1^{\mathbf{x}_i}, \cdots, p_4^{\mathbf{x}_i})$, then the loss is:

$$\mathcal{L}_{\text{detection}} = -\frac{1}{n}\sum_{i=1}^{n}\log(p_{t_i}^{\mathbf{x}_i}) \tag{5.3}$$

To make this output more human interpretable and clinically relevant we apply the following post-processing procedure:

i Convert the five output classes into three: background, benign, malignant, by summing the probabilities of both benign classes and both malignant classes.

ii Filter out background predictions.

iii Filter out foreground predictions that have a probability less than specified thresholds T_m and T_b for malignant and benign pixels, respectively.

iv Calculate contours for the remaining pixels using 4-adjacent connections to define regions.

v Calculate the convex hull for each contour, then remove those with an area less than a threshold T_{area}.

vi For each convex hull i covering an area $\mathbf{M}_i$, we assign to it a malignancy score according to the following equation:

$$c_i = \frac{\sum_{x\in\mathbf{M}_i}\mathbf{C}_x^m}{\sum\mathbf{C}_\cdot^m} \tag{5.4}$$

Where $\mathbf{C}^m$ denotes the malignant probably map. The denominator here is intended to normalize between images. This score will be used to analyze the detection results.

Examples of raw prediction results and post-processed images are shown in Figure 5.14.

Tracking System

The tracking component of our system is intended to find pixel-wise correspondence between two images of the same body part, in order to track lesions over time. This is critically important in clinical settings - rapid change in lesions is a strong indicator of cancer.

Shown in Figure 5.13, the tracking network is an adaptation of the detection network. The convolutional component of the detection network is split into two parts at `conv4_3` - the first part is preserved, and the second part's convolutions are converted into atrous convolutions. Atrous convolutions are intended to replace the use of further deconvolutional layers - they make the features finer while preserving a moderate amount of parameters [4, 80, 82].

Tracking data comes in image pairs - during the feedforward pass the two images are fed through the conv-atrous pipeline independently, after which their output is element-wise subtracted before being fed through two subsequent convolutional layers [206]. These two layers are of channel size 4096 and kernel size 7×7

and 1×1 respectively. Their output is further fed into another convolutional layer of kernel size 1×1 to output a 2D vector field of correspondences from the first to the second image of the pair - for each pixel $\mathbf{x}_i$ on one of the original images. Our model will calculate a vector shift $\mathbf{d}_i = (dx_i, dy_i)$. We use an L_2-norm loss:

$$\mathcal{L}_{\text{tracking}} = \frac{1}{n} \sum_{i=1}^{n} \|\mathbf{g}_i - \mathbf{d}_i\| \tag{5.5}$$

where $\mathbf{g}_i$ is the ground truth shift vector and n is the number of pixels that get backpropagated (note $n < M \times N$, the original image dimensions, as we only backpropagate the loss at pixels which have correspondence points - edges may not have correspondence points due to border effects of the image translation during the augmentation process).

Since the output of the tracking network has a side-length 8 times smaller than the original correspondence map, we use a Gaussian kernel to downsample the label map.

As a final step, our system searches for the best feature match within a square region of the predicted correspondence point in order to calculate the final correspondence point. We pass the input images through the network, extracting the activation features $\mathbf{f}^t$ and $\mathbf{f}^s$, for the target and source, respectively, from layer `conv4_3`. These features are bilinearly interpolated to the same size as the original image, and normalized. For a query position $\mathbf{x}_q$ in the target image, the final correspondence $\mathbf{c}^{\mathbf{x}_q}$ is calculated according to the following equation:

$$\mathbf{c}^{\mathbf{x}_q} = \arg\min_{\mathbf{x} \in \mathbf{S}} (\|\mathbf{f}^t_{\mathbf{x}_q} - \mathbf{f}^s_{\mathbf{x}}\|) + \lambda(\|\mathbf{p}_q - \mathbf{x}\|) \tag{5.6}$$

where $\mathbf{S}$ is a square of fixed side-length l (manually chosen to balance between speed and robustness) centered at the network's original correspondence prediction $\mathbf{p}_q$.

5.2.4 Experiments and Results

We use 1,300 biopsied skin lesion images and 400 high-resolution body images to generate 40000 images of size 960×960 for detection and 84000 pairs of images of size 512×512 for tracking. Of these images, 30% are held out for validation, the rest are used for training. Our system, trained on this synthetic data, is then compared to a series of baselines.

Detection

The convolutional part of our detection model (Figure 5.13) is initialized from the weights of a VGG16 network pretrained on semantic segmentation [257]. The convolutional layers in the deconvolutional part are initialized using He initialization [170] and the deconvolutional layers are initialized using a bilinear interpolation kernel. The network is trained with a fixed learning rate of $1e{-}4$. Given the redundancy in the

training set, we only train the network for 2 epochs with a batch size of 2.

Example results for detection are shown in the first three rows of Figure 5.14, whose columns show examples as they are processed from input images to raw network output to post-processed results. The parameters for post-processing T_m, T_b, and T_{area} are set to 0.85, 0.98, and 45, respectively. The hyperparameters mentioned above are chosen to maximize performance on the validation set.

Our method is compared to a baseline sliding window method (bottom row of Figure 5.14) where a 5-way classifier outputs a classification result for image patches from a test image. The baseline classifier is Google's Inception-V3 network architecture [412] pretrained on ImageNet [360] and finetuned on the 1,300 skin lesions of our dataset to output a 5-way softmax label. We extract patches at various scales from the test image using a sliding window technique over the entire image, then collect the label maps and aggregate them together into the 3-way heatmap. Afterwards, we apply the same post processing step that we use for our system in order to generate the region proposals. For comparison, we include the performance of two non-expert humans trained to detect malignant lesions and tested on the same 108 images as our detection CNN.

We test detection using 108 clinical images in which at least one skin lesion has been biopsied. For each test image, we consider the network's proposed regions, and if any region has a malignancy score greater than a threshold T, we consider it to be malignant. If the intersection-over-union (IoU) between a region and the biopsied lesion satisfies IoU > 0.5, we consider it a true positive, otherwise it is a false positive. Sweeping T in the interval [0,1] yields the receiver-operator curve shown in the left part of Figure 5.15.

As can be seen in Figure 5.14 and Figure 5.15, our method far outperforms this baseline. Our method exhibits a rapid rise in recall as the number of false positives increases, plateauing around a value of 0.8. The sliding window baseline plateaus almost immediately and does not achieve a value over 0.4. This is due in part to the fact that the sliding window baseline tends to output coarse region proposals which may contain more than one malignant lesion - in contrast, the predictions of our network tend to be highly localized to the lesions themselves. The diversity in skin backgrounds introduced by our gradient domain blending technique further helps our method to generalize better with unseen skin texture.

A human with very little training can easily detect skin lesions in general. However, analyzing the malignancy potential of each lesion is an arduous and time-consuming task, even for dermatological experts. Our detection method runs in 0.78s, and despite non-perfect performance, still has the potential to guide a doctor's eye to lesions of highest suspicion, helping to ensure that no potentially malignant lesions are missed.

To further verify potentially utility, we test two non-expert humans at the same detection task and on the same test set of images. Each human was given the 1,300 lesion images of the Edinburgh dataset, together with their benign/malignant labels. After studying the images they were then asked to detect each malignant lesion on the 108 images of our test set. As can be seen in Figure 5.15, their performance is on par with our method, but they operate near the top-right part of the curve. They exhibit reasonably good recall with a high false positive rate - as expected, they have a tendency to misclassify benign lesions as malignant.

Tracking

The convolutional and atrous sections of the tracking network are initialized with the parameters pre-trained on the detection task. He initialization [170] is used on the layers following the element-wise subtraction. The initial global learning rate is $1e{-}4$, which multiplies by a factor of 0.9 at the end of every epoch. The network is trained for 3 epochs using a batch size of 20.

For each pair of input images, the network outputs a vector field that maps one onto the other. For any position on the target image whose correspondence we want to calculate, we use the feature matching technique of equation 5.6, with $l = 64$ and $\lambda = 0.012$.

We compare our results to two baseline techniques: SIFT Flow, and Deformable Spatial Pyramids [255, 215]. We evaluate tracking accuracy using the percentage-of-correct-keypoints (PCK) metric [463]. For a given value of $\alpha \in (0,1)$, a match is considered correct if the predicted point is within $\alpha \times L$ of the ground-truth correspondence, where L is the mean diagonal length of the two images [4]. Our test set is composed of 260 pairs of correspondence labels manually annotated on temporal image pairs. These image pairs vary in pose, background, distance, viewpoint, and illumination condition.

In this particular task we do not compare to human performance, which is essentially perfect at lesion-wise correspondence matching.

Example image results are shown in Figure 5.14, where green lines show correctly predicted correspondences, and red lines show incorrect predictions. In these examples, correct predictions are those satisfying $|\mathbf{g} - \mathbf{c}| < \alpha L$, with $\alpha = 0.05$ [206], where $\mathbf{c}$ and $\mathbf{g}$ are the predicted correspondence and ground truth correspondence, respectively. We qualitatively observe superior performance. The example image pair on the left is largely dominated by a translation from one image to the next, and the image pair on the right by a difference in zoom. At this value of α our method manages to correctly match a majority of the keypoints across both pairs of images shown, achieving better performance than both baselines. SIFT Flow exhibits rather poor matching performance, likely due to the general uniformity of skin patches, for both examples. DSP underperforms our method and outperforms SIFT FLOW in the translation-dominated case, but performs worse than SIFT Flow in the zoom-dominated case on the right.

The PCK of our method, both with and without feature matching incorporated, is plotted in Figure 5.15, in comparison to both baselines. Our methods exhibit superior performance over baseline for $\alpha > 0.016$, at which point the PCK quickly climbs to 1.0 and the baselines increase linearly. This is due to the robustness of CNN features - they vary more than SIFT or DSP features on a local scale but less on a global one. For $\alpha < 0.016$ we see a steep drop off in our method's PCK, while the baselines continue to decrease in a linear fashion. This is likely due to the fact that both SIFT keypoints and the smallest spatial scales of DSP vary little as a function of image variation and thus will typically be accurately matched, ensuring the stability of both baseline curves as $\alpha \to 0$.

5.2.5 Conclusion

Here we show large-scale detection and tracking of skin lesions across images using fully convolutional neural networks in a low-data regime using domain-specific data augmentation. A key contribution of this work is a general roadmap for taking an application-domain, making it data-ready for computer vision techniques, and building a system around it. In the absence of large amounts of labeled and annotated data, we generate high volumes of synthetic data using 1,300 biopsy-proven clinical images of skin lesions and 400 body images. Skin lesion images are blended onto body images, heavily augmented with a variety of techniques, and used to train a detection network. This network is composed of a convolution component adapted from VGG16, followed by a deconvolutional component, with skip links connecting the layers of the first half to the second. We demonstrate human-interpretable detection with this method, and show that it outperforms a sliding-window baseline technique that uses a trained classifier on the same data. We then further augment the data and generate image pairs with pixel-wise correspondence between them, and use this to train a tracking network which outperforms both SIFT Flow and DSP. The tracking network's architecture is built from two weight-sharing pillars which accept each of the pair of images as input, and whose output is subtracted before being processed by subsequent layers. Each pillar is composed of part of the detection network's convolutional half (and thus initialized with those weights), and followed by atrous convolutional layers. Both networks are trained on synthetic data and tested on real-world data.

Artificial intelligence systems of this sort have the potential to improve the way healthcare is practiced, and may extend it outside of the clinic. Algorithms such as these could aid providers at spotting suspicious lesions amongst benign ones, and at observing temporal changes in lesions that may signify malignancies.

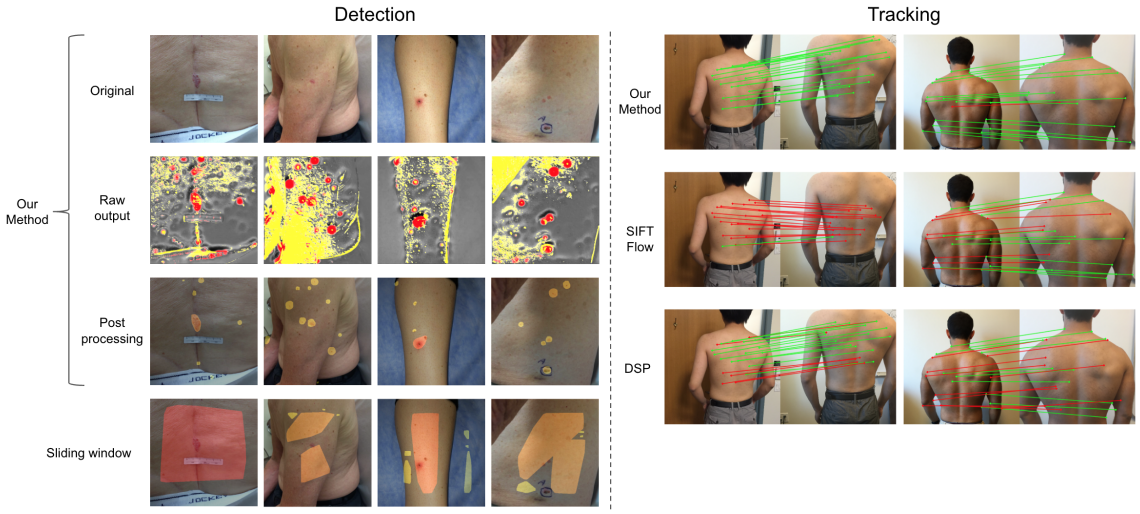


Figure 5.14: **Detection and Tracking Image Results.** Left: Detection results, Right: Tracking results. Four examples from our detection pipeline, compared to a baseline sliding-window classifier technique. Top row: original image. Second row: raw output of the network. Third row: final results after post-processing. Fourth row: final results from the baseline. Two examples from our tracking pipeline, compared to SIFT-Flow and Deformable Spatial Pyramid baselines, using $\alpha = 0.05$.

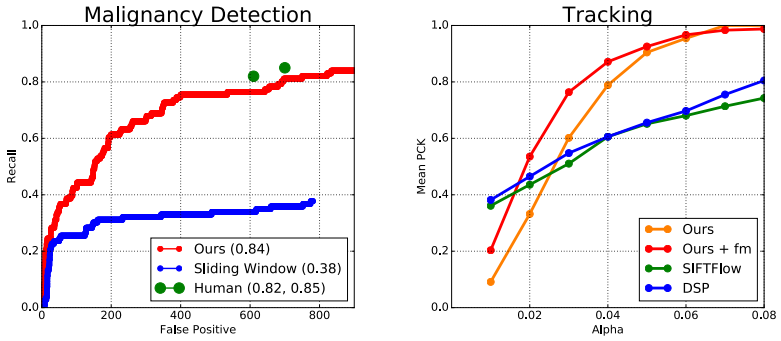


Figure 5.15: **Quantitative Results.** Detection results (top): ROC curve comparing our technique against a baseline sliding window method and two non-expert humans. Recall rate is shown in the parenthesis of the legend. Tracking Results (bottom): mean percentage of correct keypoints (PCK) as a function of $\alpha = p/L$, where p is the number of pixels, and L is the diagonal length of the image. We compare our method, with and without feature matching (fm), to SIFTFlow and Deformable Spatial Pyramids.

Bibliography

[1] Cancer Facts and Figures 2016. *American Cancer Society*, 2016.

[2] Daniel A Abrams, Srikanth Ryali, Tianwen Chen, Parag Chordia, Amirah Khouzam, Daniel J Levitin, and Vinod Menon. Inter-subject synchronization of brain responses during natural music listening. *European Journal of Neuroscience*, 37(9):1458–1469, 2013.

[3] J.K. Aggarwal and M.S. Ryoo. Human activity analysis. *ACM Computing Surveys*, 43(3):1–43, apr 2011.

[4] Pulkit Agrawal, Joao Carreira, and Jitendra Malik. Learning to see by moving. 2015.

[5] Misha B Ahrens, Michael B Orger, Drew N Robson, Jennifer M Li, and Philipp J Keller. Whole-brain functional imaging at cellular resolution using light-sheet microscopy. *Nature methods*, 10(5):413–420, 2013.

[6] Yoshihiro Akahane, Takashi Asano, Bong-Shik Song, and Susumu Noda. High-Q photonic nanocavity in a two-dimensional photonic crystal. *Nature*, 425(6961):944–947, 2003.

[7] A. Alemi-Neissi, F. B. Rosselli, and D. Zoccolan. Multifeatural shape processing in rats engaged in invariant visual object recognition. *Journal Neuroscience*, 33(14):5939–5956, 2013.

[8] A. Paul Alivisatos, Anne M. Andrews, Edward S. Boyden, Miyoung Chun, George M. Church, Karl Deisseroth, John P. Donoghue, Scott E. Fraser, Jennifer Lippincott-Schwartz, Loren L. Looger, Sotiris Masmanidis, Paul L. McEuen, Arto V. Nurmikko, Hongkun Park, Darcy S. Peterka, Clay Reid, Michael L. Roukes, Axel Scherer, Mark Schnitzer, Terrence J. Sejnowski, Kenneth L. Shepard, Doris Tsao, Gina Turrigiano, Paul S. Weiss, Chris Xu, Rafael Yuste, and Xiaowei Zhuang. Nanotools for Neuroscience and Brain Activity Mapping. *ACS Nano*, 7(3):1850–1866, 2013.

[9] A. Paul Alivisatos, Miyoung Chun, George M. Church, Ralph J. Greenspan, Michael L. Roukes, , and Rafael Yuste. The Brain Activity Map Project and the Challenge of Functional Connectomics. *Neuron*, 74, 2012.

[10] E. Aminoff, N. Gronau, and M. Bar. The parahippocampal cortex mediates spatial and nonspatial associations. *Cereb. Cortex*, 17:1493–1503, July 2007.

[11] Elissa Aminoff, Daniel L Schacter, and Moshe Bar. The cortical underpinnings of context-based memory distortion. *Journal of cognitive neuroscience*, 20(12):2226–37, December 2008.

[12] Rajagopal Ananthanarayanan, Steven K. Esser, Horst D. Simon, and Dharmendra S. Modha. The cat is out of the bag: cortical simulations with 109 neurons, 1013 synapses. In *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis*, pages 63:1–63:12, 2009.

[13] Jessica R Andrews-Hanna, Jay S Reidler, Jorge Sepulcre, Renee Poulin, and Randy L Buckner. Functional-anatomic fractionation of the brain's default network. *Neuron*, 65(4):550–62, February 2010.

[14] Jessica R Andrews-Hanna, Jonathan Smallwood, and R Nathan Spreng. The default network and self-generated thought: component processes, dynamic control, and clinical relevance. *Annals of the New York Academy of Sciences*, 1316:29–52, May 2014.

[15] Takashi Asano, Bong-Shik Song, and Susumu Noda. Analysis of the experimental q factors (˜ 1 million) of photonic crystal nanocavities. *Optics express*, 14(5):1996–2002, 2006.

[16] T. Atanasijevic, M. Shusteff, P. Fam, and A. Jasanoff. Calcium-sensitive MRI contrast agents based on superparamagnetic iron oxide nanoparticles and calmodulin. *Proceedings of the National Academy of Science*, 103(40):14707–14712, 2006.

[17] Frederico AC Azevedo, Ludmila RB Carvalho, Lea T Grinberg, José Marcelo Farfel, Renata EL Ferretti, Renata EP Leite, Roberto Lent, Suzana Herculano-Houzel, et al. Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. *Journal of Comparative Neurology*, 513(5):532–541, 2009.

[18] Alan Baddeley. The episodic buffer: a new component of working memory? *Trends in Cognitive Sciences*, 4(11):417–423, November 2000.

[19] Kelly Baker. Identity, Memory and Place. *The Word Hoard*, 1(1):Article 4, 2012.

[20] C. Baldassano, D. M. Beck, and L. Fei-Fei. Differential connectivity within the Parahippocampal Place Area. *Neuroimage*, 75:228–237, July 2013.

[21] C. Baldassano, D. M. Beck, and L. Fei-Fei. Parcellating connectivity in spatial maps. *PeerJ*, 3(e784), 2015.

[22] Lucia Ballerini, Robert B Fisher, Ben Aldridge, and Jonathan Rees. A color and texture based hierarchical K-NN approach to the classification of non-melanoma skin lesions. In *Color Medical Image Analysis*, pages 63–86. Springer, 2013.

[23] A.K. Bansal, W. Truccolo, C.E. Vargas-Irwin, and J.P. Donoghue. Decoding 3D reach and grasp from hybrid signals in motor and premotor cortices: spikes, multiunit activity, and local field potentials. *Journal of Neurophysiology*, 107(5):1337–55, 2012.

[24] Moshe Bar and Elissa Aminoff. Cortical analysis of visual context. *Neuron*, 38(2):347–58, 2003.

[25] Z. Bar-Joseph, D. K. Gifford, and T. S. Jaakkola. Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics*, 17(Suppl 1):S22–S29, jun 2001.

[26] Horace B. Barlow. Possible principles underlying the transformations of sensory messages. In W. A. Rosenblith, editor, *Sensory Communication*, pages 217–234. MIT Press, Cambridge, MA, 1961.

[27] R.P.J. Barretto and M.J. Schnitzer. In Vivo Optical Microendoscopy for Imaging Cells Lying Deep within Live Tissue. *Cold Spring Harbor Protocols*, 2012(10), 2012.

[28] G. Becker and D. Berg. Neuroimaging in basal ganglia disorders: perspectives for transcranial ultrasound. *Movement Disorders*, 16(1):23–32, 2001.

[29] Brian J Beliveau, Eric F Joyce, Nicholas Apostolopoulos, Feyza Yilmaz, Chamith Y Fonseka, Ruth B McCole, Yiming Chang, Jin Billy Li, Tharanga Niroshini Senaratne, Benjamin R Williams, et al. Versatile design and synthesis platform for visualizing genomes with oligopaint fish probes. *Proceedings of the National Academy of Sciences*, 109(52):21301–21306, 2012.

[30] C Gordon Bell, Robert Chen, and Satish Rege. Effect of technology on near term computer structures. *Computer*, 5(2):29–38, 1972.

[31] Jacob G. Bernstein and Edward S. Boyden. Optogenetic tools for analyzing the neural circuits of behavior. *Trends in Cognitive Sciences*, 15:592–600, 2011.

[32] Jacob G Bernstein, Paul A Garrity, and Edward S Boyden. Optogenetics and thermogenetics: technologies for controlling the activity of targeted cells within intact neural circuits. *Current opinion in neurobiology*, 22(1):61–71, 2012.

[33] Katherine C. Bettencourt and Yaoda Xu. The role of transverse occipital sulcus in scene perception and its relationship to object individuation in inferior intraparietal sulcus. *Journal of Cognitive Neuroscience*, 25:1711–1722, 2013.

[34] A. Bhargava and J. Gorelik. Recording single-channel currents using "smart patch-clamp" technique. *Methods Molecular Biology*, 998:189–197, 2013.

[35] A. Bhargava, X. Lin, P. Novak, K. Mehta, Y. Korchev, M. Delmar, and J. Gorelik. Super-resolution scanning patch clamp reveals clustering of functional ion channels in adult ventricular myocyte. *Circulation Ressearch*, 112(8):1112–1120, 2013.

[36] I Biederman. Recognition-by-components: a theory of human image understanding. *Psychological review*, 94(2):115–47, apr 1987.

[37] M Binder, H Kittler, A Seeber, A Steiner, H Pehamberger, and K Wolff. Epiluminescence microscopy-based classification of pigmented skin lesions using computerized image analysis and an artificial neural network. *Melanoma research*, 8(3):261–266, 1998.

[38] M Binder, H Kittler, A Seeber, A Steiner, H Pehamberger, and K Wolff. Epiluminescence microscopy-based classification of pigmented skin lesions using computerized image analysis and an artificial neural network. *Melanoma research*, 8(3):261–266, 1998.

[39] Roy Biran, Dave C Martin, and Patrick A Tresco. The brain tissue response to implanted silicon microelectrode arrays is increased when the device is tethered to the skull. *Journal of Biomedical Materials Research Part A*, 82(1):169–178, 2007.

[40] T.J. Blanche, M.A. Spacek, J.F. Hetke, and N.V. Swindale. Polytrodes: high-density silicon electrode arrays for large-scale multiunit recording. *Journal of Neurophysiology. 2005 May;93(5):. Epub 2004 Nov 17.*, 93:2987–3000, 2005.

[41] Angélique Bobrie, Marina Colombo, Graça Raposo, and Clotilde Théry. Exosome secretion: molecular mechanisms and roles in immune responses. *Traffic*, 12(12):1659–1668, 2011.

[42] Martin D Bootman, Claire Fearnley, Ioannis Smyrnias, Fraser MacDonald, and H Llewelyn Roderick. An update on nuclear calcium signalling. *Journal of Cell Science*, 122(14):2337–2350, 2009.

[43] Z Boraston and S J Blakemore. The application of eye-tracking technology in the study of autism. *The Journal of Physiology*, 581(3):893–898, 2007.

[44] Christoph Börgers, Giovanni Talei Franzesi, Fiona E. N. LeBeau, Edward S. Boyden, and Nancy J. Kopell. Minimal size of cell assemblies coordinated by gamma oscillations. *PLoS Biology*, 8(2):e1002362, 2012.

[45] U.T. Bornscheuer and M. Pohl. Improved biocatalysts by directed evolution and rational protein design. *Current Opinion Chemical Biology*, 5(2):137–43, 2001.

[46] Conrado A. Bosman, Jan-Mathijs Schoffelen, Nicolas Brunet, Robert Oostenveld, Andre M. Bastos, Thilo Womelsdorf, Birthe Rubehn, Thomas Stieglitz, Peter De Weerd, and Pascal Fries. Attentional Stimulus Selection through Selective Synchronization between Monkey Visual Areas. *Neuron*, 75:875–888, 2012.

[47] Edward S Boyden, Feng Zhang, Ernst Bamberg, Georg Nagel, and Karl Deisseroth. Millisecond-timescale, genetically targeted optical control of neural activity. *Nature Neuroscience*, 8:1263–1268, 2005.

[48] Valentino Braitenberg. *Vehicles: Experiments in Synthetic Psychology*. MIT Press, Cambridge, MA, 1986.

[49] S. Bray, A. E. Arnold, R. M. Levy, and G. Iaria. Spatial and temporal functional connectivity changes between resting and attentive states. *Hum Brain Mapp*, 36(2):549–565, 2 2015.

[50] Matthew Brett. The MNI brain and the Talairach atlas. `http://imaging.mrc-cbu.cam.ac.uk/imaging/MniTalairach`, 2002. Accessed: 2015-01-27.

[51] AA Brewer and Brian Barton. Visual field map organization in human visual cortex. In Stephane Molotchnikoff and Jean Rouat, editors, *Visual Cortex-Current Status and Perspectives*, chapter 2. 2012.

[52] Kevin L Briggman and Davi D Bock. Volume electron microscopy for neuronal circuit reconstruction. *Current Opinion in Neurobiology*, 22(1):154–161, 2012.

[53] K.L. Briggman and D.D. Bock. Current Opinion Neurobiology. *Volume electron microscopy for neuronal circuit reconstruction*, 22:154–61, 2012.

[54] K.L. Briggman, M. Helmstaedter, and W. Denk. Wiring specificity in the direction-selectivity circuit of the retina. *Nature*, 471:183–188, 2011.

[55] Andreas Buja, Deborah F Swayne, Michael L Littman, Nathaniel Dean, Heike Hofmann, and Lisha Chen. Data Visualization With Multidimensional Scaling. *Journal of Computational and Graphical Statistics*, 17(2):444–472, jun 2008.

[56] H. H. Bulthoff, S. Y. Edelman, and M. J. Tarr. How Are Three-Dimensional Objects Represented in the Brain? *Cerebral Cortex*, 5(3):247–260, may 1995.

[57] N Burgess, E a Maguire, H J Spiers, and J O'Keefe. A temporoparietal and prefrontal network for retrieving the spatial context of lifelike events. *NeuroImage*, 14(2):439–53, August 2001.

[58] Peter Burke and Christopher Rutherglen. Towards a single-chip, implantable {RFID} system: is a single-cell radio possible? *Biomedical microdevices*, 12(4):589–596, 2010.

[59] Matthew F Burkhardt, Fernando J Martinez, Sarah Wright, Carla Ramos, Dmitri Volfson, Michael Mason, Jeff Garnes, Vu Dang, Jeffery Lievers, Uzma Shoukat-Mumtaz, Rita Martinez, Hui Gai, Robert Blake, Eugeni Vaisberg, Marica Grskovic, Charles Johnson, Stefan Irion, Jessica Bright, Bonnie Cooper, Leane Nguyen, Irene Griswold-Prenner, and Ashkan Javaherian. A cellular model for sporadic ALS using patient-derived induced pluripotent stem cells. *Mol. Cell. Neurosci.*, 56:355–364, September 2013.

[60] Marco Burroni, Rosamaria Corona, Giordana DellEva, Francesco Sera, Riccardo Bono, Pietro Puddu, Roberto Perotti, Franco Nobile, Lucio Andreassi, and Pietro Rubegni. Melanoma computer-aided diagnosis. *Clinical cancer research*, 10(6):1881–1886, 2004.

[61] Stephen F. Bush. *Nanoscale Communication Networks*. Nanoscale Science and Engineering. Artech House, 2010.

[62] Daniel P. Buxhoeveden and Manuel F. Casanova. The minicolumn hypothesis in neuroscience. *Brain*, 125(5):935–951, 2002.

[63] R.B. Buxton, K. Uludag, D.J. Dubowitz, and T.T. Liu. Modeling the hemodynamic response to brain activation. *Neuroimaging*, 23:220–233, 2004.

[64] György Buzsáki. *Rhythms of the Brain*. Oxford University Press, 2006.

[65] M. J. Byrne, M. N. Waxham, and Y. Kubota. Cellular dynamic simulator: an event driven molecular simulation environment for cellular physiology. *Neuroinformatics*, 8(2):63–82, 2010.

[66] Patrick Byrne, Suzanna Becker, and Neil Burgess. Remembering the past and imagining the future: a neural model of spatial memory and imagery. *Psychological review*, 114(2):340–75, April 2007.

[67] Dong Cai, Lu Ren, Huaizhou Zhao, Chenjia Xu, Lu Zhang, Ying Yu, Hengzhi Wang, Yucheng Lan, Mary F Roberts, Jeffrey H Chuang, et al. A molecular-imprint nanosensor for ultrasensitive detection of proteins. *Nature nanotechnology*, 5(8):597–601, 2010.

[68] W. Cai, A.R. Hsu, Z. Li, and X. Chen. Are quantum dots ready for in vivo imaging in human subjects? *Nanoscale Research Letters*, 2(6):265–281, 2007.

[69] E.M. Callaway and R. Yuste. Stimulating neurons with light. *Current Opinion in Neurobiology*, 12(5):587–92, 2002.

[70] F W Campbell and J G Robson. Application of Fourier analysis to the visibility of gratings. *The Journal of physiology*, 197(3):551–66, August 1968.

[71] Ryan T. Canolty, Karunesh Ganguly, Steven W. Kennerley, Charles F. Cadieu, Kilian Koepsell, Jonathan D. Wallis, and Jose M. Carmena. Oscillatory phase coupling coordinates anatomically dispersed functional cell assemblies. *Proceedings of the National Academy of Sciences*, 107(40):17356–17361, 2010.

[72] Guan Cao, Jelena Platisa, Vincent A. Pieribone, Davide Raccuglia, Michael Kunst, and Michael N. Nitabach. Genetically targeted optical electrophysiology in intact neural circuits. *Cell*, 154:904–913, 2013.

[73] M. Carandini. From circuits to behavior: a bridge too far? *Nature Neuroscience*, 15(4):507–509, 2012.

[74] Svenja Caspers, Simon B Eickhoff, Tobias Rick, Anette Von Kapri, Torsten Kuhlen, Ruiwang Huang, Nadim J Shah, and Karl Zilles. Probabilistic fibre tract analysis of cytoarchitectonically defined human inferior parietal lobule areas reveals similarities to macaques. *NeuroImage*, 58(2):362–380, 2011.

[75] Svenja Caspers, Axel Schleicher, Mareike Bacha-Trams, Nicola Palomero-Gallagher, Katrin Amunts, and Karl Zilles. Organization of the human inferior parietal lobule based on receptor architectonics. *Cerebral Cortex*, pages 1–14, 2012.

[76] Patrick Cerwall. Ericsson mobility report. *https://www.ericsson.com/res/docs/2016/ericsson-mobility-report-2016.pdf*, pages 1–32, 2016.

[77] Patrick Cerwall. Ericsson Mobility REport: On the pulse of the networked society. 2016.

[78] Chang-Hsiao Chen, Shih-Chang Chuang, Yu-Tao Lee, Yen-Chung Chang, Shih-Rung Yeh, and Da-Jeng Yao. Three-dimensional flexible microprobe for recording the neural signal. *Journal of Micro/Nanolithography, MEMS, and MOEMS*, 9(3):031007, 2010.

[79] Chia-Chun Chen, Yen-Ping Lin, Chih-Wei Wang, Hsiao-Chien Tzeng, Chia-Hsuan Wu, Yi-Cheng Chen, Chin-Pei Chen, Li-Chyong Chen, and Yi-Chun Wu. Dna-gold nanorod conjugates for remote control of localized gene expression by near infrared irradiation. *Journal of the American Chemical Society*, 128(11):3709–3715, 2006.

[80] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *CoRR*, abs/1412.7062, 2014.

[81] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected CRFs. In *ICLR*, 2015.

[82] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *CoRR*, abs/1606.00915, 2016.

[83] Longtang L Chen, Lie-huey Lin Edward, Carol A Barnes, and Bruce L Mcnaughton. Head-direction cells in the rat posterior cortex. I . anatomical distribution and behavioral modulation. *Experimental Brain Research*, 101(1):24–34, 1994.

[84] Ting Chen and Christophe Chefdhotel. Deep learning based automatic immune cell detection for immunohistochemistry images. In Guorong Wu, Daoqiang Zhang, and Luping Zhou, editors, *Machine Learning in Medical Imaging*, Lecture Notes in Computer Science, pages 17–24. Springer International Publishing, 14 September 2014.

[85] Julong Cheng, Ali Torkamani, Yingjie Peng, Teresa M Jones, and Richard A Lerner. Plasma membrane associated transcription of cytoplasmic dna. *Proceedings of the National Academy of Sciences*, 109(27):10827–10831, 2012.

[86] KyungHyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. arXiv:1409.1259, 2014.

[87] Kwanghun Chung, Jenelle Wallace, Sung-Yon Kim, Sandhiya Kalyanasundaram, Aaron S. Andalman, Thomas J. Davidson, Julie J. Mirzabekov, Kelly A. Zalocusky, Aleksandra K. Denisin Joanna Mattis, Sally Pak, Hannah Bernstein, Charu Ramakrishnan, Logan Grosenick, Viviana Gradinaru, and Karl Deisseroth. Structural and molecular interrogation of intact biological systems. *Nature*, 10, 2013.

[88] George M. Church. Genomes for All. *Scientific American*, 294:46–54, 2006.

[89] Dan C Cireşan, Alessandro Giusti, Luca M Gambardella, and Jürgen Schmidhuber. Mitosis detection in breast cancer histology images with deep neural networks. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, pages 411–418. Springer, 2013.

[90] Wallace H Clark, David E Elder, DuPont Guerry, Leonard E Braitman, Bruce J Trock, Delray Schultz, Marie Synnestvedt, and Allan C Halpern. Model predicting survival in stage I melanoma based on tumor progression. *Journal of the National Cancer Institute*, 81(24):1893–1904, 1989.

[91] Noel Codella, Junjie Cai, Mani Abedini, Rahil Garnavi, Alan Halpern, and John R Smith. Deep learning, sparse coding, and SVM for melanoma recognition in dermoscopy images. In *International Workshop on Machine Learning in Medical Imaging*, pages 118–126, 2015.

[92] Noel Codella, Junjie Cai, Mani Abedini, Rahil Garnavi, Alan Halpern, and John R Smith. Deep learning, sparse coding, and svm for melanoma recognition in dermoscopy images. In *International Workshop on Machine Learning in Medical Imaging*, pages 118–126. Springer, 2015.

[93] Luis Pedro Coelho, Aabid Shariff, and Robert F Murphy. Nuclear segmentation in microscope cell images: a hand-segmented dataset and comparison of algorithms. In *Biomedical Imaging: From Nano to Macro, 2009. ISBI'09. IEEE International Symposium on*, pages 518–521, 2009.

[94] Jay S. Coggan, Thomas M. Bartol, Eduardo Esquenazi, Joel R. Stiles, Stephan Lamont, Maryann E. Martone, Darwin K. Berg, Mark H. Ellisman, and Terrence J. Sejnowski. Evidence for Ectopic Neurotransmission at a Neuronal Synapse. *Science*, 309(5733):446–451, 2005.

[95] Ronald A Conlon. Transgenic and gene targeted models of dementia. In *Animal Models of Dementia*, pages 77–90. Springer, 2011.

[96] R. W. Cox. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput. Biomed. Res.*, 29(3):162–173, June 1996.

[97] H. Craighead. Future lab-on-a-chip technologies for interrogating individual molecules. *Nature*, 442(7101):387–393, 2006.

[98] G Csibra and G Gergely. Social learning and social cognition: The case for pedagogy. *In Process of change in brain and cognitive development, attention and performance*, 9:152?–158, 2006.

[99] John F Davidson, Richard Fox, Dawn D Harris, Sally Lyons-Abbott, and Lawrence A Loeb. Insertion of the t3 dna polymerase thioredoxin binding domain enhances the processivity and fidelity of taq dna polymerase. *Nucleic acids research*, 31(16):4702–4709, 2003.

[100] Kevin Davies. *The $1,000 Genome: The Revolution in DNA Sequencing and the New Era of Personalized Medicine*. Free Press, New York, NY, 2010.

[101] Jacco A de Zwart, Peter van Gelderen, Xavier Golay, Vasiliki N Ikonomidou, and Jeff H Duyn. Accelerated parallel imaging for functional imaging of the human brain. *NMR in Biomedicine*, 19(3):342–351, 2006.

[102] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255, 2009.

[103] Winfried Denk and Karel Svoboda. Photon upmanship: Techreview why multiphoton imaging is more than a gimmick. *Neuron*, 18:351–357, 1997.

[104] James J. DiCarlo and David D. Cox. Untangling invariant object recognition. *Trends in Cognitive Sciences*, 11(8):333–341, 2007.

[105] P.A Dijkmans, L.J.M Juffermans, R.J.P Musters, A van Wamel, F.J ten Cate, W van Gilst, C.A Visser, N de Jong, and O Kamp. Microbubbles and ultrasound: from diagnosis to therapy. *European Journal of Echocardiography*, 5(4):245–246, 2004.

[106] Daniel A Dombeck, CD Christopher D Harvey, Lin Tian, Loren L Looger, and David W Tank. Functional imaging of hippocampal place cells at cellular resolution during virtual navigation. *Nature neuroscience*, 13(11):1433–1440, November 2010.

[107] B Dong, L Shao, M Da Costa, O Bandmann, and A F Frangi. Deep learning for automatic cell detection in wide-field microscopy zebrafish images. In *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, pages 772–776, April 2015.

[108] L. Dong, C. M. Witkowski, M. M. Craig, M. M. Greenwade, and K. L. Joseph. Cytotoxicity effects of different surfactant molecules conjugated to carbon nanotubes on human astrocytoma cells. *Nanoscale Res Lett*, 4(12):1517–1523, 2009.

[109] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[110] J. Du, T.J. Blanche, R.R. Harrison, H.A. Lester, and S.C. Masmanidis. Multiplexed, high density electrophysiology with nanofabricated neural probes. *PLoS One*, 6(10):e26204, 2011.

[111] Zhong-Wei Du, Hong Chen, Huisheng Liu, Jianfeng Lu, Kun Qian, Cindytzu-Ling Huang, Xiaofen Zhong, Frank Fan, and Su-Chun Zhang. Generation and expansion of highly pure motor neuron progenitors from human pluripotent stem cells. *Nat. Commun.*, 6:6626, 25 March 2015.

[112] K. A. Ehinger, A. Torralba, and A. Oliva. A taxonomy of visual scenes: Typicality ratings and hierarchical classification. *Journal of Vision*, 10(7):1237–1237, aug 2010.

[113] Howard Eichenbaum and Neal J. Cohen. Can We Reconcile the Declarative Memory and Spatial Navigation Views on Hippocampal Function? *Neuron*, 83(4):764–770, August 2014.

[114] Seif Eldawlatly, Rong Jin, and Karim G Oweiss. Identifying functional connectivity in large-scale neural ensemble recordings: a multiscale data mining approach. *Neural computation*, 21(2):450–477, 2009.

[115] Jeremy A Elman, Brendan I Cohn-Sheehy, and Arthur P Shimamura. Dissociable parietal regions facilitate successful retrieval of recently learned and personally familiar information. *Neuropsychologia*, 51(4):573–83, March 2013.

[116] R.A. Epstein and N. Kanwisher. A cortical representation of the local visual environment. *Nature*, 392:598–601, April 1998.

[117] Russell A Epstein, J Stephen Higgins, Karen Jablonski, and Alana M Feiler. Visual scene processing in familiar and unfamiliar environments. *Journal of neurophysiology*, 97(5):3670–83, May 2007.

[118] Russell A Epstein, Whitney E Parker, and Alana M Feiler. Where am I now? Distinct roles for parahippocampal and retrosplenial cortices in place recognition. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 27(23):6141–6149, 2007.

[119] Russell A Epstein and Lindsay K Vass. Neural systems for landmark-based wayfinding in humans. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 369(1635), 2014.

[120] Russell A. Epstein and Emily J. Ward. How reliable are visual context effects in the parahippocampal place area? *Cerebral Cortex*, 20(February):294–303, 2010.

[121] Henry Markram *et al*. The Human Brain Project: A Report to the European Commission. The HBP-PS Consortium, Lausanne, 2012.

[122] Ming Fa, Annalisa Radeghieri, Allison A Henry, and Floyd E Romesberg. Expanding the substrate repertoire of a dna polymerase by directed evolution. *Journal of the American Chemical Society*, 126(6):1748–1754, 2004.

[123] J. A. et al. Fagan. Centrifugal length separation of Carbon Nanotubes. *Langmuir*, 24:13880–13889, May 2008.

[124] Scott L Fairhall, Stefano Anzellotti, Silvia Ubaldi, and Alfonso Caramazza. Person- and place-selective neural substrates for entity-specific semantic access. *Cerebral cortex (New York, N.Y. : 1991)*, 24(7):1687–96, July 2014.

[125] Zhiguang Fan, Shan Qiao, Huang-Fu, Jiang Tao, and Li-Xin Ran. A miniaturized printed dipole antenna with v-shaped ground for 2.45 GHz RFID readers. *Progress In Electromagnetics Research*, 71:149–158, 2007.

[126] Clement Farabet, Camille Couprie, Laurent Najman, and Yann Lecun. Learning hierarchical features for scene labeling. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1915–1929, August 2013.

[127] B. Farley and W.A. Clark. Simulation of self-Organizing Systems by Digital Computer. *IRE Transactions on Information Theory*, 4:76–84, 1954.

[128] L. Fei-Fei, A. Iyer, C. Koch, and P. Perona. What do we perceive in a glance of a real-world scene? *J Vis*, 7(1):10, 2007.

[129] Li Fei-Fei, Asha Iyer, Christof Koch, and Pietro Perona. What do we perceive in a glance of a real-world scene? *Journal of Vision*, 7(1):10, jan 2007.

[130] Li Fei-Fei and Pietro Perona. A Bayesian Hierarchical Model for Learning Natural Scene Categories. 2005.

[131] Fei-Fei Fei-Fei Li and P. Perona. A Bayesian Hierarchical Model for Learning Natural Scene Categories. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 524–531. IEEE, 2005.

[132] Evan H. Feinberg, Miri K. VanHoven, Andres Bendesky, George Wang, Richard D. Fetter, Kang Shen, and Cornelia I. Bargmann. GFP Reconstitution Across Synaptic Partners (GRASP) Defines Cell Contacts and Synapses in Living Nervous Systems. *Neuron*, 57(3):353–363, 2008.

[133] Steven Finkbeiner, Michael Frumkin, and Paul D Kassner. Cell-based screening: Extracting meaning from complex data. *Neuron*, 86(1):160–174, 8 April 2015.

[134] X. et al. Gao. In vivo cancer targeting and imaging with semiconductor quantum dots. *Nature Biotechnology*, 22(8):969–976, Aug 2004.

[135] Farid J Ghadessy, Jennifer L Ong, and Philipp Holliger. Directed evolution of polymerase function by compartmentalized self-replication. *Proceedings of the National Academy of Sciences*, 98(8):4552–4557, 2001.

[136] Farid J Ghadessy, Nicola Ramsay, François Boudsocq, David Loakes, Anthony Brown, Shigenori Iwai, Alexandra Vaisman, Roger Woodgate, and Philipp Holliger. Generic expansion of the substrate spectrum of a dna polymerase by directed evolution. *Nature biotechnology*, 22(6):755–759, 2004.

[137] K.K. Ghosh, L.D. Burns, E.D. Cocker, A. Nimmerjahn, Y. Ziv, A.E. Gamal, and M.J. Schnitzer. Miniaturized integration of a fluorescence microscope. *Nature Methods*, 8(10):871–8, 2011.

[138] Gibson and James J. The perception of the visual world., 1950.

[139] James J. Gibson. *The Ecological Approach To Visual Perception*. Psychology Press, new ed edition, September 1986.

[140] V. Gilja, P. Nuyujukian, C.A. Chestek, J.P. Cunningham, B.M. Yu, J.M. Fan, M.M. Churchland, M.T. Kaufman, J.C. Kao, S.I. Ryu, and K.V. Shenoy. A high-performance neural prosthesis enabled by control algorithm design. *Nature Neuroscience*, 15:1752–1757, 2012.

[141] Ross Girshick. Fast R-CNN. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2015.

[142] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[143] Travis C Glenn. Field guide to next-generation dna sequencers. *Molecular Ecology Resources*, 11(5):759–769, 2011.

[144] James F. Glockner, Houchun H. Hu, David W. Stanley, Lisa Angelos, and Kevin King. Parallel MR imaging: A users guide. Technical Report.

[145] G Golarai, K Grill-Spector, and AL Reiss. Autism and the development of face processing. *Clinical neuroscience research*, pages 145?–160, 2006.

[146] C. Gold, D.A. Henze, and C. Koch. Using extracellular action potential recordings to constrain compartmental models. *Journal Compututational Neuroscience*, 23(1):39–58, 2007.

[147] Sally A. Goldman and Ronald Rivest. Making Maximum Entropy Computations Easier By Adding Extra Constraints. In *Proceedings of the Sixth Annual Workshop on Maximum Entropy and Bayesian Methods in Applied Statistics*, 1986.

[148] Julie D. Golomb and Nancy Kanwisher. Higher level visual cortex represents retinotopic, not spatiotopic, object location. *Cerebral Cortex*, 22(December):2794–2810, 2012.

[149] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z Ghahramani, M Welling, C Cortes, N D Lawrence, and K Q Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.

[150] Michelle R. Greene and Aude Oliva. Recognition of natural scenes from global properties: Seeing the forest without representing the trees. *Cognitive Psychology*, 58(2):137–176, 2009.

[151] Michelle R. Greene and Aude Oliva. The Briefest of Glances. *Psychological Science*, 20(4):464–472, apr 2009.

[152] Michelle R. Greene and Aude Oliva. The Briefest of Glances: The Time Course of Natural Scene Understanding. *Psychological Science*, 20(4):464–472, 2009.

[153] Michelle R. Greene and Aude Oliva. High-level aftereffects to global scene properties. *Journal of Experimental Psychology: Human Perception and Performance*, 36(6):1430–1442, dec 2010.

[154] Benjamin F. Grewe, Dominik Langer, Hansjörg Kasper, Björn M. Kampa, and Fritjof Helmchen. High-speed in vivo calcium imaging reveals neuronal network activity with near-millisecond precision. *Nature Methods*, 2:399–405, 2010.

[155] Christine Grienberger and Arthur Konnerth. Imaging Calcium in Neurons. *Neuron*, 73(5):862–885, 2012.

[156] David Gutman, Noel C F Codella, Emre Celebi, Brian Helba, Michael Marchetti, Nabin Mishra, and Allan Halpern. Skin Lesion Analysis toward Melanoma Detection: A Challenge at the International Symposium on Biomedical Imaging (ISBI) 2016, hosted by the International Skin Imaging Collaboration (ISIC). *arXiv preprint arXiv:1605.01397*, 2016.

[157] Alon Hafri, Anna Papafragou, and John C Trueswell. Getting the gist of events: recognition of two-participant actions from brief displays. *Journal of experimental psychology. General*, 142(3):880–905, aug 2013.

[158] P J Hagerman. The fragile x prevalence paradox. *Journal of medical genetics*, 45:498–?499, 2008.

[159] Scott S Hall, Michael C Frank, Guido T Pusiol, Faraz Farzin, Amy A Lightbody, and Allan L Reiss. Quantifying naturalistic social gaze in fragile x syndrome using a novel eye tracking paradigm. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 168:564?–572, 2015.

[160] Benjamin Hansen, Ying Liu, Rusen Yang, and Zhong Wang. Hybrid nanogenerator for concurrently harvesting biomechanical and biochemical energy. *ACS nano*, 4(7):3647–3652, 2010.

[161] Julia Harris, Renaud Jolivet, and David Attwell. Synaptic Energy Use and Supply. *Neuron*, 75(5):762–777, 2012.

[162] C.D. Harvey, F. Collman, D.A. Dombeck, and D.W. Tank. Intracellular dynamics of hippocampal place cells during virtual navigation. *Nature*, 461(7266):941–946, 2009.

[163] Jordan Hashemi, Thiago Vallin Spina, Mariano Tepper, Amy Esler, Vassilios Morellas, Nikolaos Papanikolopoulos, and Guillermo Sapiro. A computer vision approach for the assessment of autism-related behavioral markers. In *2012 IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL)*, pages 1–7. IEEE, 2012.

[164] Demis Hassabis, Dharshan Kumaran, and Eleanor A Maguire. Using imagination to understand the neural basis of episodic memory. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 27(52):14365–74, December 2007.

[165] Demis Hassabis and Eleanor A. Maguire. Deconstructing episodic memory with construction. *Trends in Cognitive Sciences*, 11(7):299–306, 2007.

[166] Uri Hasson, Michal Harel, Ifat Levy, and Rafael Malach. Large-scale mirror-symmetry organization of human occipito-temporal object areas. *Neuron*, 37:1027–1041, 2003.

[167] Michael Hawrylycz, Richard A. Baldock, Albert Burger, Tsutomu Hashikawa, G. Allan Johnson, Maryann E. Martone, Lydia Ng, Christopher Lau, Stephen D. Larson, Jonathan Nissanov, Luis Puelles, Seth Ruffins, Fons Verbeek, Ilya Zaslavsky, and Jyl Boline. Digital Atlasing and Standardization in the Mouse Brain. *PLoS Computational Biology*, 7, 2011.

[168] Hiroki R Hayama, Kaia L Vilberg, and Michael D Rugg. Overlap between the neural correlates of cued recall and source memory: evidence for a generic recollection network? *Journal of cognitive neuroscience*, 24(5):1127–37, May 2012.

[169] Yuichiro Hayashi, Yoshiaki Tagawa, Satoshi Yawata, Shigetada Nakanishi, and Kazuo Funabiki. Spatio-temporal control of neural activity in vivo using fluorescence microendoscopy. *European Journal of Neuroscience*, 36(6):2722–2732, 2012.

[170] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1026–1034, 2015.

[171] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[172] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. *arXiv preprint arXiv:1603. 05027*, 2016.

[173] Donald Hebb. *The Organization of Behavior*. Wiley, New York, 1949.

[174] Michael Held, Michael H A Schmitz, Bernd Fischer, Thomas Walter, Beate Neumann, Michael H Olma, Matthias Peter, Jan Ellenberg, and Daniel W Gerlich. CellCognition: time-resolved phenotype annotation in high-throughput live cell imaging. *Nat. Methods*, 7(9):747–754, 2010.

[175] Michael J Heller, Benjamin Sullivan, Dietrich Dehlinger, and Paul Swanson. Next-generation dna hybridization and self-assembly nanofabrication devices. In *Springer Handbook of Nanotechnology*, pages 389–401. Springer, 2010.

[176] F. Helmchen and W. Denk. Deep tissue two-photon microscopy. *Nat Methods*, 2(12):932–40, 2005.

[177] John M. Henderson, Christine L. Larson, and David C. Zhu. Cortical activation to indoor versus outdoor scenes: an fMRI study. *Experimental Brain Research*, 179(1):75–84, apr 2007.

[178] D. P. Hinton and C. S. Johnson, Jr. Diffusion Ordered 2D NMR Spectroscopy of Phospholipid Vesicles: Determination of Vesicle Size Distributions. *J. Phys. Chem*, 97:9064–9072, May 1993.

[179] Alan L. Hodgkin. Chance and design in electrophysiology. *The Journal of Physiology*, 263:1–21, 1976.

[180] Alan L. Hodgkin and Andrew F. Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of Physiology*, 117:500–544, 1952.

[181] J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Science*, 79(8):2554–2558, 1982.

[182] Peter R. Hoskins, Kevin Martin, and Abigail Thrush. *Diagnostic Ultrasound: Physics and Equipment*. Cambridge University Press, 2010.

[183] B. Huang, M. Bates, and X. Zhuang. Super-resolution fluorescence microscopy. *Annual Reviews Biochemistry*, 78, 2009.

[184] Bo Huang, Hazen Babcock, and Xiaowei Zhuang. Breaking the diffraction barrier: super-resolution imaging of cells. *Cell*, 143(7):1047–1058, 2010.

[185] Heng Huang, Savas Delikanli, Hao Zeng, Denise M. Ferkey, and Arnd Pralle. Remote control of ion channels and neurons through magnetic-field heating of nanoparticles. *Nature Nanonotechnology*, 5:602–606, 2010.

[186] R. S. Huang and M. I. Sereno. Bottom-up Retinotopic Organization Supports Top-down Mental Imagery. *Open Neuroimag J*, 7:58–67, 2013.

[187] Alexander G Huth, Shinji Nishimoto, An T Vu, Jack L Gallant, D. Hanlon, C.H. Anderson, K. Hatano, K. Ito, H. Fukuda, T. Schormann, and K. Zilles. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6):1210–24, dec 2012.

[188] K. Hynynen, N. McDannold, N.A. Sheikov, F.A. Jolesz, and N. Vykhodtseva. Local and reversible blood-brain barrier disruption by noninvasive focused ultrasound at frequencies suitable for trans-skull sonications. *Neuroimage*, 24(1):12–20, 2005.

[189] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[190] Marius Ctlin Iordan, Michelle R. Greene, Diane M. Beck, and Li Fei-Fei. Basic Level Category Structure Emerges Gradually across Human Ventral Visual Cortex. *Journal of Cognitive Neuroscience*, 27(7):1427–1446, jul 2015.

[191] Eugene M. Izhikevich and Gerald M. Edelman. Large-scale model of mammalian thalamo-cortical systems. *Proceedings of the National Academy of Science*, 105(9):3593–3598, 2008.

[192] Akerboom J, Chen TW, Wardill TJ, Tian L, Marvin JS, Mutlu S, Calderón NC, Esposti F, Borghuis BG, Sun XR, Gordus A, Orger MB, Portugues R, Engert F, Macklin JJ, Filosa A, Aggarwal A, Kerr RA, Takagi R, Kracun S, Shigetomi E, Khakh BS, Baier H, Lagnado L, Wang SS, Bargmann CI, Kimmel BE, Jayaraman V, Svoboda K, Kim DS, Schreiter ER, and Looger LL. Optimization of a GCaMP calcium indicator for neural activity imaging. *The Journal of Neuroscience*, 32:13819–13840, 2012.

[193] Viren Jain, H. Sebastian Seung, and Srinivas C. Turag. Machines that learn to segment images: a crucial technology for connectomics. *Current Opinion in Neurobiology*, 20(5):1–14, 2010.

[194] M. Jenkinson, C. F. Beckmann, T. E. Behrens, M. W. Woolrich, and S. M. Smith. FSL. *Neuroimage*, 62(2):782–790, August 2012.

[195] K. Jensen, J. Weldon, H. Garcia, and A. Zettl. Nanotube radio. *Nano Letters*, 7(11):3508–3511, 2007.

[196] Kaiming He Jian Sun Jifeng Dai, Yi Li. R-FCN: Object detection via region-based fully convolutional networks. *arXiv preprint arXiv:1605.06409*, 2016.

[197] John D Joannopoulos, Steven G Johnson, Joshua N Winn, and Robert D Meade. *Photonic crystals: molding the flow of light*. Princeton university press, 2011.

[198] Jeffrey D Johnson and Michael D Rugg. Recollection and the reinstatement of encoding-related cortical activity. *Cerebral cortex (New York, N.Y. : 1991)*, 17(11):2507–15, November 2007.

[199] P Jolicoeur, M A Gluck, and S M Kosslyn. Pictures and names: making the connection. *Cognitive psychology*, 16(2):243–75, apr 1984.

[200] Warren Jones and Ami Klin. Attention to eyes is present but in decline in 2-6-month-old infants later diagnosed with autism. *Nature*, 2013.

[201] Olivier R. Joubert, Guillaume A. Rousselet, Denis Fize, and Michèle Fabre-Thorpe. Processing scene context: Fast categorization and object interference. *Vision Research*, 47(26):3286–3297, 2007.

[202] Bela. Julesz. *Foundations of cyclopean perception*. MIT Press, 2006.

[203] T-P Jung, Scott Makeig, Martin J McKeown, Anthony J Bell, T-W Lee, and Terrence J Sejnowski. Imaging brain dynamics using independent component analysis. *Proceedings of the IEEE*, 89(7):1107–1122, 2001.

[204] Rafal J?zefowicz, Wojciech Zaremba, and Ilya Sutskever. An empirical exploration of recurrent network architectures. In Francis R. Bach and David M. Blei, editors, *ICML*, volume 37, pages 2342–2350, 2015.

[205] I. Kadar, O. Ben-Shahar, Ehinger K., Oliva A., and Torralba A. A perceptual paradigm and psychophysical evidence for hierarchy in scene gist processing. *Journal of Vision*, 12(13):16–16, dec 2012.

[206] Angjoo Kanazawa, David W. Jacobs, and Manmohan Chandraker. Warpnet: Weakly supervised matching for single-view reconstruction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[207] Hyo Jung Kang, Yuka Imamura Kawasawa, Feng Cheng, Ying Zhu, Xuming Xu, Mingfeng Li, Andre M. M. Sousa, Mihovil Pletikos, Kyle A. Meyer, Goran Sedmak, Tobias Guennel, Yurae Shin, Matthew B. Johnson, Zeljka Krsnik, Simone Mayer, Sofia Fertuzinhos, Sheila Umlauf, Steven N. Lisgo, Alexander Vortmeyer, Daniel R. Weinberger, Shrikant Mane, Thomas M. Hyde, Anita Huttner, Mark Reimers, Joel E. Kleinman, and Nenad Sestan. Spatio-temporal transcriptome of the human brain. *Nature*, 478:483–489, 2011.

[208] Ehud Kaplan. The M, P, and K Pathways of the Primate Visual System. In L.M. Chalupa and J.S. Werner, editors, *The Visual Neuroscience Encyclopedia*, pages 481–494. MIT Press, Cambridge, MA, 2003.

[209] John J Kasianowicz, Joseph WF Robertson, Elaine R Chan, Joseph E Reiner, and Vincent M Stanford. Nanoscopic porous sensors. *Annual Review Analytical Chemistry*, 1:737–766, 2008.

[210] Kendrick N. Kay, Thomas Naselaris, Ryan J. Prenger, and Jack L. Gallant. Identifying natural images from human brain activity. *Nature*, 452(7185):352–355, mar 2008.

[211] K.N. Kay, T. Naselaris, R.J. Prenger, and J.L. Gallant. Identifying natural images from human brain activity. *Nature*, 452:352–355, 2008.

[212] K. Kempa, J. Rybczynski, Z. Huang, K. Gregorczyk, A. Vidan, B. Kimball, J. Carlson, G. Benham, Y. Wang, A. Herczynski, and Z.???F. Ren. Carbon Nanotubes as Optical Antennae. *Advanced Materials*, 19(3):421–426, 2007.

[213] R. A. Kerr, T. M. Bartol, B. Kaminsky, M. Dittrich, J. C. Chang, S. B. Baden, T. J. Sejnowski, and J. R. Stiles. Fast Monte Carlo Simulation Methods for Biological Reaction-Diffusion Systems in Solution and on Surfaces. *SIAM Journal Scientific Computing*, 30(6):3126, 2008.

[214] Hongkeun Kim. Dissociating the roles of the default-mode, dorsal, and ventral networks in episodic memory retrieval. *NeuroImage*, 50(4):1648–57, May 2010.

[215] Jaechul Kim, Ce Liu, Fei Sha, and Kristen Grauman. Deformable spatial pyramid matching for fast dense correspondences. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013.

[216] P. Kim, M. Puoris'haag, D. Co Te, C.P. Lin, and S.H. Yun. In vivo confocal and multiphoton microendoscopy. *Journal of Biomedical Optics*, 13(1):010501, 2008.

[217] Danielle R King, Marianne De Chastelaine, Rachael L Elward, Tracy H Wang, and Michael D Rugg. Recollection-Related Increases in Functional Connectivity Predict Individual Differences in Memory Accuracy. *The Journal of Neuroscience*, 35(4):1763–1772, 2015.

[218] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412. 6980*, 2014.

[219] Harold Kittler, H Pehamberger, K Wolff, and M Binder. Diagnostic accuracy of dermoscopy. *The lancet oncology*, 3(3):159–165, 2002.

[220] Ami Klin, Warren Jones, Robert Schultz, Fred Volkmar, and Donald Cohen. Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism. *Archives of general psychiatry*, 59(9):809–816, 2002.

[221] C. Kocabas, H. Kim, T. Banks, J.A. Rogers, A.A. Pesetski, J.E. Baumgardner, S.V. Krishnaswamy, and H. Zhang. Radio frequency analog electronics based on carbon nanotube transistors. *Proceedings of the National Academy of Science*, 105(5):1405–1409, 2008.

[222] Suhasa B. Kodandaramaiah, Giovanni T. Franzesi, Brian Y. Chow, Edward S. Boyden, and Craig R. Forest. Automated whole-cell patch-clamp electrophysiology of neurons in vivo. *Nature Methods*, 9(6):585–587, 2012.

[223] Hartmuth C. Kolb, M. G. Finn, and K. Barry Sharpless. Click Chemistry: Diverse Chemical Function from a Few Good Reactions. *Angewandte Chemie International Edition*, 40(11):2004–2021, 2001.

[224] Julia Kollewe. DNA machine can sequence human genomes in hours. *The Guardian*, Feb 2012.

[225] Ravi K Komanduri, Chulwoo Oh, and Michael J Escuti. Reflective liquid crystal polarization gratings with high efficiency and small pitch. *Proceedings SPIE 7050, Liquid Crystals XII*, pages 70500J–70500J, 2008.

[226] Talia Konkle and Alfonso Caramazza. Tripartite organization of the ventral stream by animacy and object size. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 33(25):10235–42, 2013.

[227] J.G. Koomey, S. Berard, M. Sanchez, and H. Wong. Implications of historical trends in the electrical efficiency of computing. *Annals of the History of Computing, IEEE*, 33(3):46–54, 2011.

[228] Konrad Kording. Of toasters and molecular ticker tapes. *PLoS Computational Biology*, 7(12):e1002291, 2011.

[229] J. Korlach, P.J. Marks, R.L. Cicero, J.J. Gray, D.L. Murphy, D.B. Roitman, T.T. Pham, G.A. Otto, M. Foquet, and S.W. Turner. Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures. *Proceedings of the National Academy of Science*, 105(4):1176–1181, 2008.

[230] J.M. Kralj, A.D. Douglass, D.R. Hochbaum, D. Maclaurin, and A.E. Cohen. Optical recording of action potentials in mammalian neurons using a microbial rhodopsin. *Nature Methods*, 9(1):90–5, 2012.

[231] Dwight J Kravitz, Kadharbatcha S Saleem, Chris I Baker, and Mortimer Mishkin. A new neural framework for visuospatial processing. *Nature Reviews Neuroscience*, 12(4):217–230, 2011.

[232] Dwight J Kravitz, Kadharbatcha S Saleem, Chris I Baker, Leslie G Ungerleider, and Mortimer Mishkin. The ventral visual pathway: an expanded neural framework for the processing of object quality. *Trends in cognitive sciences*, 17(1):26–49, January 2013.

[233] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[234] Brice A Kuhl and Marvin M Chun. Successful remembering elicits event-specific activity patterns in lateral parietal cortex. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 34(23):8051–60, June 2014.

[235] Yatindra Kumar, M L Dewal, and R S Anand. Epileptic seizure detection using DWT based fuzzy approximate entropy and support vector machine. *Neurocomputing*, 133, June 2014.

[236] André Kurs, Aristeidis Karalis, Robert Moffatt, J Joannopoulos, Peter Fisher, and Marin Soljacic. Wireless power transfer via strongly coupled magnetic resonances. *Science (New York, N.Y.)*, 317(5834):83–86, 2007.

[237] Takaaki Kuwajima, Austen A. Sitko, Punita Bhansali, Chris Jurgens, William Guido, and Carol Mason. ClearT: a detergent- and solvent-free clearing method for neuronal and non-neuronal tissue. *Development*, 140:1364–1368, 2013.

[238] Kestutis Kveraga, Avniel Singh Ghuman, Karim S Kassam, Elissa a Aminoff, Matti S Hämäläinen, Maximilien Chaumon, and Moshe Bar. Early onset of neural synchronization in the contextual associations network. *Proceedings of the National Academy of Sciences of the United States of America*, 108:3389–3394, 2011.

[239] M. Kyoung and E. D. Sheets. Vesicle diffusion close to a membrane: intermembrane interactions measured with fluorescence correlation spectroscopy. *Biophys. J.*, 95(12):5789–5797, Dec 2008.

[240] Thomas K. Landauer and Susan T. Dumais. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240, 1997.

[241] S.B. Laughlin, R.R. de Ruyter van Steveninck, and J.C. Anderson. The metabolic cost of neural information. *Nature Neuroscience*, 1(1):36–41, 1998.

[242] S. Lazebnik, C. Schmid, and J. Ponce. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 2161–2168, Washington, DC, USA, 2006. IEEE Computer Society.

[243] S. Lazebnik, C. Schmid, and J. Ponce. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2 (CVPR'06)*, volume 2, pages 2169–2178. IEEE, 2006.

[244] K.N. Le. *Orthogonal Frequency Division Multiplexing with Diversity for Future Wireless Systems*. Bentham Science Publishers, 2012.

[245] J. Lecoq, P. Tiret, M. Najac, G.M. Shepherd, C.A. Greer, and S. Charpak. Odor-evoked oxygen consumption by action potential and synaptic transmission in the olfactory bulb. *Journal Neuroscience*, 29(5):1424–1433, 2009.

[246] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[247] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[248] Yann LeCun, Yoshua Bengio, and others. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

[249] Yulia Lerner, Christopher J Honey, Lauren J Silbert, and Uri Hasson. Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 31(8):2906–15, February 2011.

[250] Li Li, Roger J. Zemp, Gina Lungu, George Stoica, and Lihong V. Wang. Photoacoustic imaging of lacz gene expression in vivo. *Journal of Biomedical Optics*, 12(2):020504–020504–3, 2007.

[251] Meng-Lin Li, Jung-Taek Oh, X. Xie, G. Ku, Wei Wang, Chun Li, G. Lungu, G. Stoica, and L.V. Wang. Simultaneous molecular and hypoxia imaging of brain tumors in vivo using spectroscopic photoacoustic tomography. *Proceedings of the IEEE*, 96(3):481–489, 2008.

[252] Yunzhu Li, Benyuan Sun, Tianfu Wu, and Yizhou Wang. Face detection with end-to-end integration of a convnet and a 3d model. In *ECCV*, 2016.

[253] Drew Linsley and Sean P Macevoy. Encoding-Stage Crosstalk Between Object- and Spatial Property-Based Scene Processing Pathways. *Cerebral cortex (New York, N.Y. : 1991)*, March 2014.

[254] A. M. Litke. The retinal readout system: an application of microstrip detector technology to neurobiology. *Nuclear Instruments and Methods in Physics Research Section A*, 418:203–209, 1998.

[255] Ce Liu, Jenny Yuen, Antonio Torralba, Josef Sivic, and William T Freeman. Sift flow: Dense correspondence across different scenes. In *European conference on computer vision*, pages 28–42. Springer, 2008.

[256] Po-Ru Loh, Michael Baym, and Bonnie Berger. Compressive genomics. *Nature biotechnology*, 30(7):627–630, 2012.

[257] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *CVPR*, November 2015.

[258] Jonathan L Long, Ning Zhang, and Trevor Darrell. Do convnets learn correspondence? In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1601–1609. Curran Associates, Inc., 2014.

[259] Xi Long, W Louis Cleveland, and Y Lawrence Yao. Multiclass detection of cells in multicontrast composite images. *Comput. Biol. Med.*, 40(2):168–178, 2010.

[260] Lester C. Loschky and Adam M. Larson. The natural/man-made distinction is made before basic-level distinctions in scene gist processing. *Visual Cognition*, 18(4):513–536, apr 2010.

[261] A. Y. Louie, M. M. Huber, E. T. Ahrens, U. Rothbacher, R. Moats, R. E. Jacobs, S. E. Fraser, and T. J. Meade. In vivo visualization of gene expression using magnetic resonance imaging. *Nature Biotechnology*, 18(3):321–325, 2000.

[262] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[263] Bruno Madore, P. Jason White, Kai Thomenius, and Gregory T. Clement. Accelerated Focused Ultrasound Imaging. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 56(12):2612–2623, 2009.

[264] Reinoud Maex and Erik De Schutter. Resonant synchronization in heterogeneous networks of inhibitory neurons. *The Journal of neuroscience*, 23(33):10503–10514, 2003.

[265] P.J. Magistretti and L. Pellerin. Cellular mechanisms of brain energy metabolism and their relevance to functional brain imaging. *Philosophyical Transactions Royal Society London B Biological Science*, 354(1387):1155–1163, 1999.

[266] J. T. Mannion and H. G. Craighead. Nanofluidic structures for single biomolecule fluorescent detection. *Biopolymers*, 85(2):131–143, 2007.

[267] Yunxiang Mao, Zhaozheng Yin, and Joseph M Schober. Iteratively training classifiers for circulating tumor cell detection. In *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, pages 190–194. IEEE, 2015.

[268] Adam Marblestone. Class discussion. Personal Communication, 2013.

[269] Adam H. Marblestone, Bradley M. Zamft, Yael G. Maguire, Mikhail G. Shapiro, Thaddeus R. Cybulski, Joshua I. Glaser, Ben Stranges, Reza Kalhor, Elad Alon David A. Dalrymple, Dongjin Seo, Michel M. Maharbiz, Jose Carmena, Jan Rabaey, Edward S. Boyden, George M. Church, and Konrad P. Kording. Physical principles for scalable neural recording. *ArXiv preprint cs.CV/1306.5709*, 2013.

[270] Steven A Marchette, Lindsay K Vass, Jack Ryan, and Russell A Epstein. Anchoring the neural compass: coding of local spatial reference frames in human medial parietal lobe. *Nature neuroscience*, October 2014.

[271] Henry Markram. The Blue Brain Project. *Nature Reviews Neuroscience*, 7:153–160, 2006.

[272] Henry Markram, Joachim Lübke, Michael Frotscher, and Bert Sakmann. Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science*, 275(5297):213–215, 1997.

[273] D Marr and WH Vision. Freeman and Company. *New York*, 1982.

[274] Jesse D. Marshall and Mark J. Schnitzer. Optical Strategies for Sensing Neuronal Voltage Using Quantum Dots and Other Semiconductor Nanocrystals. *ACS Nano*, 7:4601–4609, 2013.

[275] Ammara Masood and Adel Ali Al-Jumaily. Computer aided diagnostic support system for skin cancer: a review of techniques and algorithms. *International journal of biomedical imaging*, 2013, 2013.

[276] Sonal Mazumder, Rajib Dey, M.K. Mitra, S. Mukherjee, and G.C. Das. Review: Biofunctionalized Quantum Dots in Biology and Medicine. *Journal of Nanomaterials*, 2009, 2009.

[277] W. S. McCulloch and W. H. Pitts. A Logical Calculus of Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics*, 5:115–133, 1943.

[278] A. Medina, I. Matta, and J. Byers. On the origin of power laws in Internet topologies. *Computer Communication Review*, 30:2:18–28, 2000.

[279] S W Menzies, L M Bischof, G Peden, H G Talbot, A Gutenev, R L Thompson, K W McNamara, G Burlutski, W H McCarthy, and V N Skladnev. Automated instrumentation for the diagnosis of invasive melanoma: image analysis of oil epiluminescence microscopy. In *Skin Cancer and UV Radiation*, pages 1064–1070. Springer, 1997.

[280] Salma Mesmoudi, Vincent Perlbarg, David Rudrauf, Arnaud Messe, Basile Pinsard, Dominique Hasboun, Claudia Cioli, Guillaume Marrelec, Roberto Toro, Habib Benali, and Yves Burnod. Resting state networks' corticotopy: the dual intertwined rings architecture. *PloS one*, 8(7):e67444, January 2013.

[281] A.A. Mezer, J. Yeatman, N. Stikov, K.N. Kay, N-J. Cho, R.F. Dougherty, M.L. Perry, J. Parvizi, L.H. Hua, K. Butts-Pauly, and B.A. Wandell. Measuring within the voxel: Brain macromolecular tissue volume in individual subjects. *Nature Medicine (in press)*, 2013.

[282] Luana Micallef and Peter Rodgers. eulerAPE: drawing area-proportional 3-Venn diagrams using ellipses. *PloS one*, 9(7):e101717, 2014.

[283] Kristina D. Micheva and Stephen J. Smith. Array Tomography: A New Tool for Imaging the Molecular Architecture and Ultrastructure of Neural Circuits. *Neuron*, 55(1):25–36, 2007.

[284] Shawn Mikula, Jonas Binding, and Winfried Denk. Staining and embedding the whole mouse brain for electron microscopy. *Nature Methods*, 9:1198–1201, 2012.

[285] George A. Miller and George A. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, nov 1995.

[286] Seung Kyu Min, Woo Youn Kim, Yeonchoo Cho, and Kwang S Kim. Fast dna sequencing with a graphene-based nanochannel device. *Nature nanotechnology*, 6(3):162–165, 2011.

[287] Y. Mishchenko. Automation of 3D reconstruction of neural tissue from large volume of conventional serial section transmission electron micrographs. *Journal Neuroscience Methods*, 176(2):276–89, 2009.

[288] Mortimer Mishkin, Leslie G. Ungerleider, and Kathleen A. Macko. Object vision and spatial vision: two cortical pathways. *Trends in Neurosciences*, 6:414–417, January 1983.

[289] T. M. Mitchell, S. V. Shinkareva, A. Carlson, K.M. Chang, V. L. Malave, R. A. Mason, and M. A. Just. Predicting Human Brain Activity Associated with the Meanings of Nouns. *Science*, 320(5880):1191–1195, 2008.

[290] R.D. Mitra, J. Shendure, J. Olejnik, E-K. Olejnik, and Church G.M. Fluorescent in situ sequencing on polymerase colonies. *Analytical Biochemistry*, 320(1):55–65, 2003.

[291] A. Miyawaki, J. Llopis, R. Heim, J. M. McCaffery, J. A. Adams, M. Ikura, and R. Y. Tsien. Fluorescent indicators for Ca2+ based on green fluorescent proteins and calmodulin. *Nature*, 388(6645):882–887, Aug 1997.

[292] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, and others. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[293] Daniela Montaldi, Tom J Spencer, Neil Roberts, and Andrew R Mayes. The neural system that mediates familiarity memory. *Hippocampus*, 16(5):504–20, January 2006.

[294] Gordon E Moore et al. Cramming more components onto integrated circuits, 1965.

[295] Leonard Mordfin, editor. *Handbook of Reference Data for Nondestructive Testing*. American Society for Testing & Materials, West Conshohocken, PA, 2002.

[296] Vernon B Mountcastle. Modality and topographic properties of single neurons of cats somatic sensory cortex. *J. neurophysiol*, 20(4):408–434, 1957.

[297] Vernon B. Mountcastle. The columnar organization of the neocortex. *Brain*, 120(4):701–722, 1997.

[298] Vernon B. Mountcastle. Introduction to the Special Issue on Computation in Cortical Columns. *Cerebral Cortex*, 13(1):2–4, January 2003.

[299] Eran A Mukamel, Axel Nimmerjahn, and Mark J Schnitzer. Automated analysis of cellular signals from large-scale calcium imaging data. *Neuron*, 63(6):747–760, 2009.

[300] K. B. Mullis. The unusual origin of the polymerase chain reaction. *Scientific American*, 262(4):56–61, 1990.

[301] Herbert Muschamp. The Secret History of 2 Columbus Circle, January 2006.

[302] J.L. Nadeau, S.J. Clarke, C.A. Hollmann, and D.M. Bahcheli. Quantum dot-FRET systems for imaging of neuronal action potentials. *IEEE Engineering in Medicine & Biology Society Conference*, 1, 2006.

[303] K Nakamura, R Kawashima, N Sato, a Nakamura, M Sugiura, T Kato, K Hatano, K Ito, H Fukuda, T Schormann, and K Zilles. Functional delineation of the human occipito-temporal areas related to face and scene processing. A PET study. *Brain : a journal of neurology*, 123 (Pt 9:1903–1912, 2000.

[304] S. Nasr, N. Liu, K. J. Devaney, X. Yue, R. Rajimehr, L. G. Ungerleider, and R. B. Tootell. Scene-selective cortical regions in human and nonhuman primates. *J. Neurosci.*, 31(39):13771–13785, September 2011.

[305] Shahin Nasr, Kathryn J Devaney, and Roger B H Tootell. Spatial encoding and underlying circuitry in scene-selective cortex. *NeuroImage*, 83:892–900, December 2013.

[306] Shahin Nasr, Cesar E Echavarria, and Roger B H Tootell. Thinking outside the box: rectilinear shapes selectively activate scene-selective cortex. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 34(20):6721–35, May 2014.

[307] David E. Newman-Toker, Ali S. Saber Tehrani, Georgios Mantokoudis, John H. Pula, Cynthia I. Guede, Kevin A. Kerber, Ari Blitz, Sarah H. Ying, Yu-Hsiang Hsieh, Richard E. Rothman, Daniel F. Hanley, David S. Zee, and Jorge C. Kattah. Quantitative Video-Oculography to Help Diagnose Stroke in Acute Vertigo and Dizziness: Toward an ECG for the Eyes. *Stroke*, 44(4):1158–1161, 2013.

[308] Dwight Nishimura. Discussion concerning MRI technology. Personal Comminication, 2013.

[309] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1520–1528, 2015.

[310] Kengo Nozaki, Takasumi Tanabe, Akihiko Shinya, Shinji Matsuo, Tomonari Sato, Hideaki Taniyama, and Masaya Notomi. Sub-femtojoule all-optical switching using a photonic-crystal nanocavity. *Nature Photonics*, 4(7):477–483, 2010.

[311] K M O'Craven and N Kanwisher. Mental imagery of faces and places activates corresponding stimulus-specific brain regions. *Journal of cognitive neuroscience*, 12:1013–1023, 2000.

[312] S Ogawa, T M Lee, A R Kay, and D W Tank. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciences*, 87(24):9868–9872, 1990.

[313] A. Oliva and P. G. Schyns. Diagnostic colors mediate scene recognition. *Cogn Psychol*, 41(2):176–210, September 2000.

[314] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42:145–175, 2001.

[315] Aude Oliva and Antonio Torralba. Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[316] B. A. Olshausen and D. J. Field. Natural image statistics and efficient coding. *Computation in Neural Systems*, 7(2):333–339, 1996.

[317] Hassana Oyibo, Gang Cao, Huiqing Zhan, Alex Koulakov, Lynn Enquist, Joshua Dubnau, and Anthony Zador. Converting neural circuit connectivity into a high-throughput dna sequencing problem using pseudorabies virus. Unpublished, 2013.

[318] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

[319] B.S. Paratala, B.D. Jacobson, S. Kanakia, L.D. Francis, and B. Sitharaman. Physicochemical characterization, and relaxometry studies of micro-graphite oxide, graphene nanoplatelets, and nanoribbons. *PLoS One*, 7(6):e38185, 2012.

[320] S. Park, T. Konkle, and A. Oliva. Parametric Coding of the Size and Clutter of Natural Scenes in the Human Brain. *Cereb. Cortex*, January 2014.

[321] Soojin Park and Marvin M Chun. Different roles of the parahippocampal place area (PPA) and retrosplenial cortex (RSC) in panoramic scene perception. *NeuroImage*, 47(4):1747–56, October 2009.

[322] B. N. Pasley, S. V. David, N. Mesgarani, A. Flinker, S. A. Shamma, N. E. Crone, R. T. Knight, and E. F. Chang. Reconstructing speech from human auditory cortex. *PLoS Biology*, 10(1):e1001251, 2012.

[323] G. Patterson and J. Hays. SUN attribute database: Discovering, annotating, and recognizing scene attributes. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2751–2758. IEEE, jun 2012.

[324] Genevieve Patterson, Chen Xu, Hang Su, and James Hays. The SUN Attribute Database: Beyond Categories for Deeper Scene Understanding. *International Journal of Computer Vision*, 108(1-2):59–81, may 2014.

[325] Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. WordNet::Similarity: measuring the relatedness of concepts, 2004.

[326] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. *ACM Trans. Graph.*, 22(3):313–318, July 2003.

[327] A. Perron, W. Akemann, H. Mutoh, and T. Knopfel. Genetically encoded probes for optical imaging of brain electrical activity. *Progress Brain Research*, 196:63–77, 2012.

[328] Darcy Peterka, Hiroto Takahashi, and Rafael Yuste. Imaging voltage in neurons. *Neuron*, 69(1):9–21, 2011.

[329] Jan Peters, Irene Daum, Elke Gizewski, Michael Forsting, and Boris Suchan. Associations evoked during memory encoding recruit the context-network. *Hippocampus*, 19(2):141–51, February 2009.

[330] D. Pivonka, A. Yakovlev, A.S.Y. Poon, and T. Meng. A millimeter-sized wirelessly powered and remotely controlled locomotive implant. *Biomedical Circuits and Systems, IEEE Transactions on*, 6(6):523–532, 2012.

[331] Vadim S Polikov, Patrick A Tresco, and William M Reichert. Response of brain tissue to chronically implanted neural electrodes. *Journal of neuroscience methods*, 148(1):1–18, 2005.

[332] M C Potter. Short-term conceptual memory for pictures. *Journal of experimental psychology. Human learning and memory*, 2(5):509–22, sep 1976.

[333] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 413–420. IEEE, jun 2009.

[334] T. Ragan, L. R. Kadiri, K. U. Venkataraju, K. Bahlmann, J. Sutin, J. Taranda, I. Arganda-Carreras, Y. Kim, H. S. Seung, and P. Osten. Serial two-photon tomography for automated ex vivo mouse brain imaging. *Nature Methods*, 9(3):255–258, 2012.

[335] R. Rajimehr, K. J. Devaney, N. Y. Bilenko, J. C. Young, and R. B. Tootell. The "parahippocampal place area" responds preferentially to high spatial frequencies in humans and monkeys. *PLoS Biol.*, 9(4):e1000608, April 2011.

[336] V. S. Ramachandran. Perception of shape from shading. *Nature*, 331(6152):163–166, jan 1988.

[337] Kiran Ramlakhan and Yi Shang. A mobile automated skin lesion classification system. In *Tools with Artificial Intelligence (ICTAI), 2011 23rd IEEE International Conference on*, pages 138–141, 2011.

[338] Kiran Ramlakhan and Yi Shang. A mobile automated skin lesion classification system. In *2011 IEEE 23rd International Conference on Tools with Artificial Intelligence*, pages 138–141. IEEE, 2011.

[339] Bharath Ramsundar, Steven Kearnes, Patrick Riley, Dale Webster, David Konerding, and Vijay Pande. Massively multitask networks for drug discovery. *arXiv preprint arXiv:1502. 02072*, 2015.

[340] Charan Ranganath and Maureen Ritchey. Two cortical systems for memory-guided behaviour. *Nature reviews. Neuroscience*, 13(10):713–26, October 2012.

[341] R. P. N. Rao, B. A. Olshausen, and M. S. Lewicki. *Probabilistic Models of the Brain: Perception and Neural Function*. MIT Press, Cambridge, MA, 2002.

[342] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. CNN Features off-the-shelf: an Astounding Baseline for Recognition. mar 2014.

[343] James M. Rehg, Agata Rozga, Gregory D. Abowd, and Matthew S. Goodwin. Behavioral imaging and autism. *IEEE Pervasive Computing*, 13(2):84–87, 2014.

[344] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Neural Information Processing Systems (NIPS)*, 2015.

[345] Laura Walker Renninger and Jitendra Malik. When is scene identification just texture recognition? *Vision Research*, 44(19):2301–2311, sep 2004.

[346] Ronald A. Rensink. Change Detection. *Annual Review of Psychology*, 53(1):245–277, feb 2002.

[347] Juan F Restrepo, Gastón Schlotthauer, and María E Torres. Maximum approximate entropy and r threshold: A new approach for regularity changes detection. arXiv:1405.7637(nlin.CD), 2014.

[348] M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nat. Neurosci.*, 2(11):1019–1025, November 1999.

[349] Alessandra Rigamonti, Giuliana G Repetti, Chicheng Sun, Feodor D Price, Danielle C Reny, Francesca Rapino, Karen Weisinger, Chen Benkler, Quinn P Peterson, Lance S Davidow, Emil M Hansson, and Lee L Rubin. Large-Scale production of mature neurons from human pluripotent stem cells in a Three-Dimensional suspension culture system. *Stem Cell Reports*, 6(6):993–1008, 14 June 2016.

[350] G. A. Robertson. High-resolution scanning patch clamp: life on the nanosurface. *Circulation Ressearch*, 112(8):1088–1090, 2013.

[351] Howard W Rogers, Martin A Weinstock, Steven R Feldman, and Brett M Coldiron. Incidence estimate of nonmelanoma skin cancer (keratinocyte carcinomas) in the us population, 2012. *JAMA dermatology*, 151(10):1081–1086, 2015.

[352] Howard W. Rogers, Martin A. Weinstock, Steven R. Feldman, Brett M. Coldiron, Wang YG, and Tang JY. Incidence Estimate of Nonmelanoma Skin Cancer (Keratinocyte Carcinomas) in the US Population, 2012. *JAMA Dermatology*, 151(10):1081, 10 2015.

[353] A Roggan, M Friebel, K Do Rschel, A Hahn, and G Mu Ller. Optical properties of circulating human blood in the wavelength range 400-2500 nm. *Journal of biomedical optics*, 4(1):36–46, 1999.

[354] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241, 2015.

[355] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015.

[356] Barbara Rosado, Scott Menzies, Alexandra Harbauer, Hubert Pehamberger, Klaus Wolff, Michael Binder, and Harald Kittler. Accuracy of computer diagnosis of melanoma: a quantitative meta-analysis. *Archives of Dermatology*, 139(3):361–367, 2003.

[357] Eleanor Rosch, Carolyn B Mervis, Wayne D Gray, David M Johnson, and Penny Boyes-Braem. Basic objects in natural categories. *Cognitive Psychology*, 8(3):382–439, 1976.

[358] C.J. Rozell, D.H Johnson, R.G. Baraniuk, and B.A. Olshausen. Sparse coding via thresholding and local competition in neural circuits. *Neural Computation*, 20(10):2526–2563, 2008.

[359] Douglas B Rusch, Aaron L Halpern, Granger Sutton, Karla B Heidelberg, Shannon Williamson, Shibu Yooseph, Dongying Wu, Jonathan A Eisen, Jeff M Hoffman, Karin Remington, Karen Beeson, Bao Tran, Hamilton Smith, Holly Baden-Tillson, Clare Stewart, Joyce Thorpe, Jason Freeman, Cynthia Andrews-Pfannkoch, Joseph E Venter, Kelvin Li, Saul Kravitz, John F Heidelberg, Terry Utterback, Yu-Hui Rogers, Luisa I Falcon, Valeria Souza, Germain Bonilla-Rosso, Luis E Eguiarte, David M Karl, Shubha Sathyendranath, Trevor Platt, Eldredge Bermingham, Victor Gallardo, Giselle Tamayo-Castillo, Michael R Ferrari, Robert L Strausberg, Kenneth Nealson, Robert Friedman, Marvin Frazier, and J. Craig Venter. The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol*, 5(3):e77, 2007.

[360] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[361] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, and others. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[362] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: A database and web-based tool for image annotation, 2008.

[363] Bryan Russell, Antonio Torralba, Kevin Murphy, and William Freeman. LabelMe: A Database and Web-Based Tool for Image Annotation. *International Journal Computer Vision*, 77(1-3):157–173, 2008.

[364] A. Sadek. Class discussion. Personal Communication, 2013.

[365] A.S. Sadek, R.B. Karabalin, J. Du, M.L. Roukes, C. Koch, and S.C. Masmanidis. Wiring Nanoscale Biosensors with Piezoelectric Nanomechanical Resonators. *Nano Letters*, 10:1769–1773, 2010.

[366] Murat Saglam, Yuki Hayashida, and Nobuki Murayama. A Retinal Circuit Model Accounting for Functions of Amacrine Cells. In Masumi Ishikawa, Kenji Doya, Hiroyuki Miyamoto, and Takeshi Yamakawa, editors, *Neural Information Processing*, pages 1–6. Springer-Verlag, Berlin, Heidelberg, 2008.

[367] Gabriel Salerni, Cristina Carrera, Louise Lovatto, Rosa M Martí-Laborda, Guillermina Isern, Josep Palou, Llúcia Alós, Susana Puig, and Josep Malvehy. Characterization of 1152 lesions excised over 10 years using total-body photography and digital dermatoscopy in the surveillance of patients at high risk for melanoma. *Journal of the American Academy of Dermatology*, 67(5):836–845, 2012.

[368] G. Salimi-Khorshidi, G. Douaud, C. F. Beckmann, M. F. Glasser, L. Griffanti, and S. M. Smith. Automatic denoising of functional MRI data: combining independent component analysis and hierarchical fusion of classifiers. *Neuroimage*, 90:449–468, April 2014.

[369] Uzma Samadani and David Heeger. Discussions regarding the prospects for neurological assessment using eye-trackng technology. Personal Communication, 2013.

[370] Massimo Scanziani and Michael Hausser. Electrophysiology in the age of light. *Nature*, 461(7266):930–939, 2009.

[371] R. Schalek, N. Kasthuri, K. Hayworth, D. Berger, J. Tapia, J. Morgan, S. Turaga, E. Fagerholm, H. Seung, and J. Lichtman. Development of high-throughput, high-resolution 3D reconstruction of large-volume biological tissue using automated tape collection ultramicrotomy and scanning electron microscopy. *Microscopic Microanalysis*, 1752:966–967, 2011.

[372] D. Scheinost, T. Stoica, J. Saksa, X. Papademetris, R.T. Constable, C. Pittenger, and M. Hampson. Orbitofrontal cortex neurofeedback produces lasting changes in contamination anxiety and resting-state connectivity. *Translational Psychiatry*, 3:e250, 2013.

[373] L. Schermelleh, R. Heintzmann, and H. Leonhardt. A guide to super-resolution fluorescence microscopy. *Journal of Cell Biology*, 190:165–175, 2010.

[374] T Schindewolf, Wilhelm Stolz, Rene Albert, Wolfgang Abmayr, and Harry Harms. Classification of melanocytic lesions with color and texture analysis using digital image processing. *Analytical and quantitative cytology and histology/the International Academy of Cytology [and] American Society of Cytology*, 15(1):1–11, 1993.

[375] Andreas Schindler and Andreas Bartels. Parietal cortex codes for egocentric space beyond the field of view. *Current Biology*, 23(2):177–82, January 2013.

[376] F Schroff, D Kalenichenko, and J Philbin. Facenet: A unified embedding for face recognition and clustering. *Proc. IEEE*, 2015.

[377] Lori Schuh and Ivo Drury. Intraoperative electrocorticography and direct cortical electrical stimulation. *Seminars in Anesthesia, Perioperative Medicine and Pain*, 16:46–55, 1997.

[378] Jon A. Schuller, Edward S. Barnard, Wenshan Cai, Young Chul Jun, Justin S. White, and Mark L. Brongersma. Plasmonics for extreme light concentration and manipulation. *Nature Materials*, 9:193–204, 2010.

[379] K Seidl, S Spieth, S Herwik, J Steigert, R Zengerle, O Paul, and P Ruther. In-plane silicon probes for simultaneous neural recording and drug delivery. *Journal of Micromechanics and Microengineering*, 20(10):105006, 2010.

[380] Dongjin Seo, Jose M. Carmena, Jan M. Rabaey, Elad Alon, , and Michel M. Maharbiz. Neural dust: An ultrasonic, low power solution for chronic brain-machine interfaces. *ArXiv preprint cs.CV/1307.2196*, 2013.

[381] Anne B. Sereno and Sidney R. Lehky. Population coding of visual space: comparison of spatial representations in the dorsal and ventral pathways. *Frontiers in Computational Neuroscience*, 4(0):159, 2010.

[382] Pierre Sermanet, David Eigen, Xiang Zhang, Michael Mathieu, Rob Fergus, and Yann LeCun. OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. dec 2013.

[383] Pierre Sermanet, David Eigen, Xiang Zhang, Michael Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *International Conference on Learning Representations (ICLR 2014)*. CBLS, April 2014.

[384] Sebastian Seung. *Connectome: How the Brain's Wiring Makes Us Who We Are*. Houghton Mifflin Harcourt, Boston, 2012.

[385] M. G. Shapiro, G. G. Westmeyer, P. A. Romero, J. O. Szablowski, B. Kuster, A. Shah, C. R. Otey, R. Langer, F. H. Arnold, and A. Jasanoff. Directed evolution of a magnetic resonance imaging contrast agent for noninvasive imaging of dopamine. *Nature Biotechnololgy*, 28(3):264–270, 2010.

[386] Chen Change Loy Shuo Yang, Ping Luo and Xiaoou Tang. From facial parts responses to face detection: A deep learning approach. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[387] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, and others. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[388] R.H. Silverman. High-resolution ultrasound imaging of the eye — a review. *Clinical Experimental Ophthalmology*, 37(1):54–67, 2009.

[389] Patrice Y Simard, David Steinkraus, and John C Platt. Best practices for convolutional neural networks applied to visual document analysis. In *ICDAR*, volume 3, pages 958–962, 2003.

[390] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[391] Ronak Singhal. Inside intel next generation nehalem microarchitecture. In *Hot Chips*, volume 20, 2008.

[392] S. M. Smith, A. Hyvarinen, G. Varoquaux, K. L. Miller, and C. F. Beckmann. Group-PCA for very large fMRI datasets. *Neuroimage*, 101:738–749, November 2014.

[393] H J Spiers and E a Maguire. The neuroscience of remote spatial memory: a tale of two cities. *Neuroscience*, 149(1):7–27, October 2007.

[394] O. Sporns, G. Tononi, and R. Kotter. The human connectome: A structural description of the human brain. *PLoS Computataional Biology*, 1(4):e42, 2005.

[395] RN Spreng, RA Mar, and ASN Kim. The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: a quantitative meta-analysis. *Journal of cognitive neuroscience*, 7:489–510, 2009.

[396] Sarah A Stanley, Jennifer E Gagner, Shadi Damanpour, Mitsukuni Yoshida, Jonathan S Dordick, and Jeffrey M Friedman. Radio-wave heating of iron oxide nanoparticles can regulate plasma glucose in mice. *Science Signaling*, 336(6081):604, 2012.

[397] Dustin E. Stansbury, Thomas Naselaris, and Jack L. Gallant. Natural Scene Statistics Account for the Representation of Scene Categories in Human Visual Cortex. *Neuron*, 79(5):1025–1034, sep 2013.

[398] Richard B Stein, E Roderich Gossen, and Kelvin E Jones. Neuronal variability: noise or part of the signal? *Nature Reviews Neuroscience*, 6(5):389–397, 2005.

[399] G.S. Stent, W.B. Kristan, W.O. Friesen, C.A. Ort, M. Poon, and R.L. Calabrese. Neuronal generation of the leech swimming movement. *Science*, 200(4348):1348–1357, 1978.

[400] Robert S Stern. Prevalence of a history of skin cancer in 2007: results of an incidence-based model. *Archives of dermatology*, 146(3):279–282, 2010.

[401] Robert S. Stern, Youl P, Green A, MacLennan R, Martin NG, Margolis DJ, and Stern RS. Prevalence of a History of Skin Cancer in 2007. *Archives of Dermatology*, 146(3):313–317, 3 2010.

[402] Joel R. Stiles, Thomas M. Bartol, Miriam M. Salpeter, and Terrence J. Sejnowski Edwin E. Salpeter. Synaptic Variability: New Insights from Reconstructions and Monte Carlo Simulations with MCell Synapses. In W. Maxwell Cowan, Thomas C. Sudhof, and Charles F. Stevens, editors, *Synapses*, pages 681–731. Johns Hopkins University Press, Baltimore, MD, 2001.

[403] Alexandros Stougiannis, Farhan Tauheed, Mirjana Pavlovic, Thomas Heinis, and Anastasia Ailamak. Data-driven Neuroscience: Enabling Breakthroughs Via Innovative Data Management. In *ACM SIGMOD International Conference on Management of Data*. ACM, 2013.

[404] M. R. Stratton, P. J. Campbell, and P. A. Futreal. The cancer genome. *Nature*, 458:719–724, 2009.

[405] Steven H. Strogatz. *Nonlinear Dynamics And Chaos: With Applications To Physics, Biology, Chemistry, And Engineering*. Wiley, New York, 2002.

[406] Hao Su, Charles R. Qi, Yangyan Li, and Leonidas J. Guibas. Render for cnn: Viewpoint estimation in images using cnns trained with rendered 3d model views. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[407] Kelly Sullivan, Deborah D Hatton, Julie Hammer, John Sideris, Stephen Hooper, Peter A Ornstein, and Donald B Bailey. Sustained attention and response inhibition in boys with fragile X syndrome: measures of continuous performance. *American Journal of Medical Genetics. Part B: Neuropsychiatric Genetics*, 144B(4):517–532, 2007.

[408] Susan Sunkin, Lydia Ng, Christopher Lau, Tim Dolbeare, Terri L. Gilbert, Carol L. Thompson, Michael Hawrylycz, and Chinh Dang. Allen Brain Atlas: an integrated spatio-temporal portal for exploring the central nervous system. *Nucleic Acids Research*, 41:996–1008, 2013.

[409] Christian Szegedy, Sergey Ioffe, and Vincent Vanhoucke. Inception-v4, Inception-ResNet and the impact of residual connections on learning. *CoRR*, abs/1602.07261, 2016.

[410] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[411] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

[412] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *arXiv preprint arXiv:1512.00567*, 2015.

[413] Karl K Szpunar, Jason C K Chan, and Kathleen B McDermott. Contextual processing in episodic future thought. *Cerebral cortex (New York, N.Y. : 1991)*, 19(7):1539–48, July 2009.

[414] Karl K Szpunar, Peggy L St Jacques, Clifford A Robbins, Gagan S Wig, and Daniel L Schacter. Repetition-related reductions in neural activity reveal component processes of mental simulation. *Social cognitive and affective neuroscience*, 9(5):712–22, May 2014.

[415] Martin Szummer and Rosalind W Picard. Indoor-Outdoor Image Classiication. 1998.

[416] A Takashima, K M Petersson, F Rutters, I Tendolkar, O Jensen, M J Zwarts, B L McNaughton, and G Fernández. Declarative memory consolidation in humans: a prospective functional magnetic resonance imaging study. *Proceedings of the National Academy of Sciences of the United States of America*, 103(3):756–61, January 2006.

[417] F. Tauheed, L. Biveinis, T. Heinis, F. Schurmann, H. Markram, and A. Ailamaki. Accelerating Range Queries for Brain Simulations. In *IEEE International Conference on Data Engineering*, pages 941–952. IEEE Computer Society, 2012.

[418] Jonas Thelin, Henrik Jörntell, Elia Psouni, Martin Garwicz, Jens Schouenborg, Nils Danielsen, and Cecilia Eriksson Linsmeier. Implant size and fixation mode strongly influence tissue reactions in the cns. *PloS one*, 6(1):e16267, 2011.

[419] B. Titze and W. Denk. Automated in-chamber specimen coating for serial block-face electron microscopy. *Journal Microscopy*, 250(2):101–110, 2013.

[420] A. Torralba, R. Fergus, and W.T. Freeman. 80 Million Tiny Images: A Large Data Set for Nonparametric Object and Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(11):1958–1970, nov 2008.

[421] Antonio Torralba, Rob Fergus, and Yair Weiss. Small Codes and Large Image Databases for Recognition. In *Proceedings of IEEE Computer Vision and Pattern Recognition*, pages 1–8. IEEE Computer Society, 2008.

[422] W. Truccolo, G.M. Friehs, J.P. Donoghue, and L.R. Hochberg. Primary motor cortex tuning to intended movement kinematics in humans with tetraplegia. *Journal Neurosci*, 28(5):1163–78, 2008.

[423] X. Tu, S. Manohar, A. Jagota, and M. Zheng. DNA sequence motifs for structure-specific recognition and separation of carbon nanotubes. *Nature*, 460(7252):250–253, Jul 2009.

[424] R.S. Tucker and K. Hinton. Energy consumption and energy density in optical and electronic signal processing. *Photonics Journal, IEEE*, 3(5):821–833, 2011.

[425] Barbara Tversky and Kathleen Hemenway. Categories of environmental scenes. *Cognitive Psychology*, 15(1):121–149, 1983.

[426] Mitsouko van Assche, Valeria Kebets, Patrik Vuilleumier, and Frederic Assal. Functional Dissociations Within Posterior Parietal Cortex During Scene Integration and Viewpoint Changes. *Cereb Cortex*, page bhu215, September 2014.

[427] M. van Buuren, M. C. W. Kroes, I. C. Wagner, L. Genzel, R. G. M. Morris, and G. Fernandez. Initial Investigation of the Effects of an Experimentally Learned Schema on Spatial Associative Memory in Humans. *Journal of Neuroscience*, 34(50):16662–16670, December 2014.

[428] Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. 25 January 2016.

[429] D. C. Van Essen, S. M. Smith, D. M. Barch, T. E. Behrens, E. Yacoub, K. Ugurbil, and WU-Minn HCP Consortium. The WU-Minn Human Connectome Project: an overview. *Neuroimage*, 80:62–79, October 2013.

[430] David A Van Valen, Takamasa Kudo, Keara M Lane, Derek N Macklin, Nicolas T Quach, Mialy M DeFelice, Inbal Maayan, Yu Tanouchi, Euan A Ashley, and Markus W Covert. Deep learning automates the quantitative analysis of individual cells in Live-Cell imaging experiments. *PLoS Comput. Biol.*, 12(11):e1005177, November 2016.

[431] O.J. van Vlijmen, F.A. Rangel, S.J. Berge, E.M. Bronkhorst, A.G. Becking, and Kuijpers-Jagtman A.M. Measurements on 3D models of human skulls derived from two different cone beam CT scanners. *Clinical Oral Investigation*, 15:721–727, 2011.

[432] Seralynne D Vann, John P Aggleton, and E A Maguire. What does the retrosplenial cortex do? *Nature Reviews Neuroscience*, 10(11):792–802, November 2009.

[433] Lindsay K Vass and Russell A Epstein. Abstract representations of location and facing direction in the human brain. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 33(14):6133–42, April 2013.

[434] Kimberly Venta, Gabriel Shemer, Matthew Puster, Julio A. Rodriguez-Manzo, Adrian Balan, Jacob K. Rosenstein, Ken Shepard, and Marija Drndic. Differentiation of short, single-stranded dna homopolymers in solid-state nanopores. *ACS Nano*, 7(5):4629–4636, 2013.

[435] Ricardo Vigário, Jaakko Sarela, V Jousmiki, Matti Hamalainen, and Erkki Oja. Independent component approach to the analysis of EEG and MEG recordings. *Biomedical Engineering, IEEE Transactions on*, 47(5):589–593, 2000.

[436] Kaia L Vilberg and Michael D Rugg. Memory retrieval and the parietal cortex: a review of evidence from a dual-process perspective. *Neuropsychologia*, 46(7):1787–99, January 2008.

[437] Kaia L Vilberg and Michael D Rugg. Functional significance of retrieval-related activity in lateral parietal cortex: Evidence from fMRI and ERPs. *Human brain mapping*, 30(5):1490–501, May 2009.

[438] Kaia L Vilberg and Michael D Rugg. The neural correlates of recollection: transient versus sustained FMRI effects. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 32(45):15679–87, November 2012.

[439] Oriol Vinyals, Ł ukasz Kaiser, Terry Koo, Slav Petrov, Ilya Sutskever, and Geoffrey Hinton. In *Advances in Neural Information Processing Systems 28*, pages 2773–2781.

[440] Julia Vogel and Bernt Schiele. Semantic Modeling of Natural Scenes for Content-Based Image Retrieval. *International Journal of Computer Vision*, 72(2):133–157, apr 2007.

[441] Joshua T. Vogelstein, Brendon O. Watson, Adam M. Packer, Rafael Yuste, Bruno Jedynak, and Liam Paninski. Spike Inference from Calcium Imaging Using Sequential Monte Carlo Methods. *Biophysical Journal*, 97:636–655, 2009.

[442] M.M. Waldrop. Computer modelling: Brain in a box. *Nature*, 482(7386):456–8, 2012.

[443] William Grey Walter. An imitation of life. *Scientific American*, 182(5):42–45, 1950.

[444] B.A. Wandell and J.D. Yeatman. Biological development of reading circuits. *Current Opinion Neurobiology*, 23:261–268, 2013.

[445] Brian Wandell. Class discussion concerning MRI technology. Personal Communication, 2013.

[446] Brian A. Wandell, Serge O. Dumoulin, and Alyssa A. Brewer. Visual Field Maps in Human Cortex. *Neuron*, 56(2):366–383, 2007.

[447] L. Wang, R. E. B. Mruczek, M. J. Arcaro, and S. Kastner. Probabilistic maps of visual topography in human cortex. *Cerebral Cortex*, 2014.

[448] Lihong V. Wang. Prospects of photoacoustic tomography. *Medical Physics*, 35(12):5758–5767, 2008.

[449] Lihong V. Wang and Song Hu. Photoacoustic Tomography: In Vivo Imaging from Organelles to Organs. *Science*, 335(6075):1458–1462, 2012.

[450] Po-Hsun Wang, Hao-Li Liu, Po-Hung Hsu, Chia-Yu Lin, Churng-Ren Chris Wang, Pin-Yuan Chen, Kuo-Chen Wei, Tzu-Chen Yen, and Meng-Lin Li. Gold-nanorod contrast-enhanced photoacoustic micro-imaging of focused-ultrasound induced blood-brain-barrier opening in a rat model. *Journal of Biomedical Optics*, 17:061222, 2012.

[451] X. Wang, Y. Pang, G. Ku, X. Xie, G. Stoica, and L. V. Wang. Noninvasive laser-induced photoacoustic tomography for structural and functional in vivo imaging of the brain. *Nature Biotechnology*, 21(7):803–806, 2003.

[452] Emily J Ward, Sean P MacEvoy, and Russell A Epstein. Eye-centered encoding of visual space in scene-selective regions. *Journal of vision*, 10:6, 2010.

[453] D.B. Weibel, P. Garstecki, D. Ryan, W.R. DiLuzio, M. Mayer, J.E. Seto, and G.M. Whitesides. Microoxen: microorganisms to move microscale loads. *Proceedings of the National Academy of Science*, 102(34):11963–11967, 2005.

[454] Alison J. Wiggett and Paul E. Downing. Representation of Action in Occipito-temporal Cortex. *Journal of Cognitive Neuroscience*, 23(7):1765–1780, jul 2011.

[455] Brian A Wilt, Laurie D Burns, Eric Tatt Wei Ho, Kunal K Ghosh, Eran A Mukamel, and Mark J Schnitzer. Advances in light microscopy for neuroscience. *Annual review of neuroscience*, 32:435, 2009.

[456] L Wittgenstein. *Philosophical investigations*. 2010.

[457] Y. Wu, J.A. Phillips, H. Liu, R. Yang, and W. Tan. Carbon nanotubes protect DNA strands during cellular delivery. *ACS Nano*, 2(10):2023–2028, 2008.

[458] Gang Xia, Liangjing Chen, Takashi Sera, Ming Fa, Peter G Schultz, and Floyd E Romesberg. Directed evolution of novel polymerase activities: mutation of a dna polymerase into an efficient rna polymerase. *Proceedings of the National Academy of Sciences*, 99(10):6597–6602, 2002.

[459] J Xiao, KA Ehinger, J Hays, and A Torralba. Sun database: Exploring a large collection of scene categories. *International Journal of*, 2016.

[460] Jianxiong Xiao, Krista A. Ehinger, James Hays, Antonio Torralba, and Aude Oliva. Sun database: Exploring a large collection of scene categories. *International Journal of Computer Vision*, 2014.

[461] Yan Xu, Yang Li, Mingyuan Liu, Yipei Wang, Maode Lai, Eric I Chang, and Others. Gland instance segmentation by deep multichannel side supervision. *arXiv preprint arXiv:1607. 03222*, 2016.

[462] Xinmai Yang, Erich W. Stein, S. Ashkenazi, and Lihong V. Wang. Nanoparticles for photoacoustic imaging. *Wiley Interdisciplinary Reviews: Nanomedicine and Nanobiotechnology*, 1(4):360–368, 2009.

[463] Yi Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, pages 1385–1392, Washington, DC, USA, 2011. IEEE Computer Society.

[464] Bangpeng Yao and Li Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 17–24. IEEE, jun 2010.

[465] A Yaroslavsky, P Schulze, I Yaroslavsky, R Schober, F Ulrich, and H Schwarzmaier. Optical properties of selected native and coagulated human brain tissues in vitro in the visible and near infrared spectral range. *Physics in medicine and biology*, 47(12):2059–2073, 2002.

[466] Frances A. Yates. *The Art of Memory*. University of Chicago Press, Chicago, IL, 1966.

[467] B. T. T. Yeo, F. M. Krienen, S. B. Eickhoff, S. N. Yaakub, P. T. Fox, R. L. Buckner, C. L. Asplund, and M. W. L. Chee. Functional Specialization and Flexibility in Human Association Cortex. *Cerebral Cortex*, page bhu217, September 2014.

[468] B T Thomas Yeo, Fenna M Krienen, Jorge Sepulcre, Mert R Sabuncu, Danial Lashkari, Marisa Hollinshead, Joshua L Roffman, Jordan W Smoller, Lilla Zöllei, Jonathan R Polimeni, Bruce Fischl, Hesheng Liu, and Randy L Buckner. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of neurophysiology*, 106(3):1125–65, September 2011.

[469] O. Yizhar, L.E. Fenno, T.J. Davidson, M. Mogri, and K. Deisseroth. Optogenetics in neural systems. *Neuron*, 71(1):9–34, 2011.

[470] Anthony Zador. Class discussion. Personal Communication, 2013.

[471] Anthony M. Zador, Joshua Dubnau, Hassana K. Oyibo, Huiqing Zhan, Gang Cao, and Ian D. Peikon. Sequencing the Connectome. *PLoS Biology*, 10(10):e1001411, 2012.

[472] Kareem A. Zaghloul and Kwabena Boahen. Visual Prosthesis and Ophthalmic Devices: New Hope in Sight. In Joyce Tombran-Tink and Joseph F. Rizzo III Colin J. Barnstable and, editors, *Circuit Designs That Model the Properties of the Outer and Inner Retina*, pages 135–158. Springer-Verlag, Berlin, Heidelberg, 2007.

[473] Bradley Michael Zamft, Adam H. Marblestone, Konrad Kording, Daniel Schmidt, Daniel Martin-Alarcon, Keith Tyo, Edward S. Boyden, and George Church. Measuring Cation Dependent DNA Polymerase Fidelity Landscapes by Deep Sequencing. *PLoS ONE*, 7(8):e43876, 2012.

[474] F. Zhang, V. Gradinaru, A.R. Adamantidis, R. Durand, R.D. Airan, L. de Lecea, and K. Deisseroth. Optogenetic interrogation of neural circuits: technology for probing mammalian brain structures. *Nature Protocols*, 5(3):439–56, 2010.

[475] Nan Zhang, Xiaodi Su, Paul Free, Xiaodong Zhou, Koon Gee Neoh, Jinghua Teng, and Wolfgang Knoll. Plasmonic metal nanostructure array by glancing angle deposition for biosensing application. *Sensors and Actuators B: Chemical*, 183(0):310–318, 2013.

[476] Y. Zhang and T. Wang. Quantum Dot Enabled Molecular Sensing and Diagnostics. *Theranostics*, 2(7):631–654, 2012.

[477] Qing Zhong, Alberto Giovanni Busetto, Juan P Fededa, Joachim M Buhmann, and Daniel W Gerlich. Unsupervised modeling of cell morphology dynamics for time-lapse microscopy. *Nat. Methods*, 9(7):711–713, 2012.

[478] Tinghui Zhou, Philipp Krahenbuhl, Mathieu Aubry, Qixing Huang, and Alexei A. Efros. Learning dense correspondence via 3d-guided cycle consistency. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[479] Xiangxin Zhu and Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, pages 2879–2886. IEEE, 2012.

[480] Yaniv Ziv, Laurie D Burns, Eric D Cocker, Elizabeth O Hamel, Kunal K Ghosh, Lacey J Kitch, Abbas El Gamal, and Mark J Schnitzer. Long-term dynamics of CA1 hippocampal place codes. *Nature neuroscience*, 2013.

[481] D. Zoccolan, N. Oertelt, J. J. DiCarlo, and D. D. Cox. A rodent model for the study of invariant visual object recognition. *Proceedings of the National Academy of Sciences*, 106(21):8748–8753, 2009.

[482] A. N. Zorzos, J. Scholvin, E. S. Boyden, and C. G. Fonstad. Three-dimensional multiwaveguide probe array for light delivery to distributed brain circuits. *Optics Letters*, 23:4841–4843, 2012.