

Statistical and Deep Learning Frameworks for High Throughput Neuronal Signal and  
Image Processing

by

Somayyeh Soltanian-Zadeh

Department of Biomedical Engineering  
Duke University

Date: \_\_\_\_\_

Approved:

\_\_\_\_\_  
Sina Farsiu, Advisor

\_\_\_\_\_  
Yiyang Gong

\_\_\_\_\_  
Joseph A. Izatt

\_\_\_\_\_  
Warren M. Grill

\_\_\_\_\_  
Junjie Yao

Dissertation submitted in partial fulfillment of  
the requirements for the degree of Doctor  
of Philosophy in the Department of  
Biomedical Engineering in the Graduate School  
of Duke University

2020

ABSTRACT

Statistical and Deep Learning Frameworks for High Throughput Neuronal Signal and  
Image Processing

by

Somayyeh Soltanian-Zadeh

Department of Biomedical Engineering  
Duke University

Date: \_\_\_\_\_

Approved:

\_\_\_\_\_  
Sina Farsiu, Advisor

\_\_\_\_\_  
Yiyang Gong

\_\_\_\_\_  
Joseph A. Izatt

\_\_\_\_\_  
Warren M. Grill

\_\_\_\_\_  
Junjie Yao

An abstract of a dissertation submitted in partial  
fulfillment of the requirements for the degree  
of Doctor of Philosophy in the Department of  
Biomedical Engineering in the Graduate School of  
Duke University

2020

Copyright by  
Somayyeh Soltanian-Zadeh  
2020

## Abstract

Quantitative analysis of the central nervous system (CNS) - comprised of the brain, the spinal cord, and the eyes - for a deeper intuition into its function often requires *in vivo* visualization of its microscopic structures. For the brain, calcium imaging using genetically encoded calcium indicators (GECIs) allows targeted, large-scale imaging of neuronal populations with cellular resolution in animals. Combined with closed-loop optogenetic control of single cells, neuroscientists can potentially test population-based models of the underlying neuronal system, adding a significant body of knowledge to the field. Realization of a closed-loop optical neuronal control system currently lacks computational frameworks (e.g., neuron segmentation) that drive the system's components based on recent data. Current neuron segmentation methods either require the acquisition of the full movie or are unable to reliably identify active neurons.

On another front, *in vivo* visualization of retinal cells has become possible with the incorporation of adaptive optics (AO) into existing retinal imaging systems, such as optical coherence tomography (OCT). A complete morphometric analysis of the living human retina at cellular level could potentially improve diagnosis, treatment planning, and monitoring of retinal diseases. The current standard approach for quantifying ganglion cells (GCs; one of the fundamental cell types for vision) from AO-OCT volumes is manual, making the task highly subjective, time consuming, and thus not feasible for large-scale studies and clinical use.

This dissertation describes the development of computational frameworks for accurate analysis of neurons from high-resolution optical images of the brain and the retina. In part 1, a statistical and information theoretic framework was developed for quantifying the resolution limit and the Cramer Rao lower bound (CRB) in detecting closely timed neuronal spikes from two-photon calcium imaging recordings. Monte-Carlo simulations with biologically derived parameters were used to numerically calculate the resolution limit and compare the performance of the optimal estimators with the CRB. Additionally, we applied our detector to distinguish overlapping transients from experimentally obtained calcium imaging data.

In part 2, a fast and robust framework was developed to automatically segment active neurons from two-photon calcium imaging recordings. A convolutional neural network (CNN) is at the core of the framework which exploits the spatiotemporal information in the recorded movies. The method is validated using two separate online datasets and its performance is compared against other state-of-the-art techniques.

In part 3, the focus is shifted to analyzing AO-OCT images of the human retina. We developed a weakly-supervised deep learning-based method to automatically segment GCs in the AO-OCT volumetric images. We validated the performance of our framework using images from healthy and glaucoma subjects acquired with two different imagers across various retinal locations and compared the performance with expert graders.

In conclusion, this dissertation provides a set of statistical and deep learning frameworks for high throughput neuronal signal and image processing. Our modern computational frameworks can be used to rapidly and accurately parse neuronal activity from calcium imaging data and measure neuronal biomarkers for *in vivo* monitoring of retinal diseases. The presented automatic frameworks have comparable performance to human experts in detecting brain neurons and retinal GCs, which is important in the long-term goal to facilitate the monitoring of microscopic structures of the CNS through optimal quantification tools.

## **Dedication**

To my wonderful parents Nahid and Hamid, my incredible sisters Sepeedah and Sameeraa, and my beloved husband Ehsan.

# Contents

Abstract.....	iv
List of Tables.....	xii
List of Figures.....	xiii
Acknowledgements.....	xv
1. Introduction.....	1
1.1 Motivation.....	1
1.2 Innovation.....	5
2. Fundamental Limits in Resolving Neural Spikes in Two-photon Calcium Imaging Recordings.....	7
2.1 Introduction.....	7
2.2 Methods.....	10
2.2.1 Action Potential Evoked Fluorescence Signal Model.....	10
2.2.2 Statistical Analysis of Resolution.....	11
2.2.3 Extraction of Single and Double AP Evoked Fluorescence Transients.....	15
2.2.4 The Cramer-Rao Lower Bound.....	18
2.3 Results.....	20
2.3.1 The Gamma Distribution Characterizes the Peak Amplitude.....	21
2.3.2 The Detector Distinguishes Two Fluorescence Transients with ISI on the Order of Tens of Milliseconds.....	23
2.3.3 Prior Knowledge about Signal Amplitudes Yields Theoretically Equal ISI Estimation Performance to the Known Case.....	28

2.3.4 Maximum Likelihood and Maximum a Posteriori Estimators Closely Approach the Theoretical Bounds .....	32
2.4 Discussion .....	35
3. Active Neuron Segmentation from Two-Photon Calcium Imaging Recordings Using Deep Learning .....	39
3.1 Introduction .....	40
3.2 Methods .....	44
3.2.1 Allen Brain Observatory Dataset and Labeling .....	44
3.2.2 Neurofinder Dataset and Labeling .....	47
3.2.3 Proposed Active Neuron Segmentation Method .....	48
3.2.4 Image Processing Steps .....	50
3.2.5 Neural Network Architecture .....	50
3.2.6 Training the Network .....	52
3.2.7 Post-processing .....	53
3.2.8 Spike Detection and Discriminability Index .....	55
3.2.9 Quantification of Peak Signal-to-noise Ratio (PSNR) .....	57
3.2.10 Evaluation Metrics .....	57
3.2.11 Speed Analysis .....	58
3.2.12 Quantification and Statistical Analysis .....	59
3.2.13 Hardware .....	59
3.3 Results .....	59
3.3.1 STNeuroNet Accurately Segmented Neurons from the Allen Brain Observatory .....	59

3.3.2 Trained STNeuroNet Segmented Neurons from Unseen Recordings of Additional Cortical Layers .....	67
3.3.3 STNeuroNet Accurately Segmented Neurons from the Neurofinder Dataset .....	69
3.4 Discussion .....	73
4. Instance Segmentation of Ganglion Cells from AO-OCT Volumes via Weakly-Supervised Deep Learning .....	81
4.1 Introduction .....	81
4.2 Methods .....	86
4.2.1 AO-OCT Datasets .....	87
4.2.2 Ophthalmic Examination and Glaucoma Diagnosis .....	88
4.2.3 Study Design .....	88
4.2.4 GCL Soma Segmentation .....	90
4.2.5 Performance Evaluation .....	97
4.3 Results .....	98
4.3.1 Achieving Expert Performance on Healthy Subjects and Generalizing to an Unseen Retinal Location .....	98
4.3.2 Achieving Expert Performance on Glaucoma Patients .....	104
4.3.3 Structural and Functional Characteristics of Glaucomatous Eyes Differ from Control Eyes .....	109
4.3.4 Generalizing Between Imaging Devices .....	112
4.4 Discussion .....	113
5. Conclusion .....	118
Appendix A: Calculation of the Fisher Information Matrix .....	120

Appendix B: Other Algorithms for Automatic Neuron Segmentation from Two-photon Calcium imaging Videos.....	122
B.1 CaImAn.....	122
B.2 Suite2p .....	124
B.3 HNCcorr .....	125
B.4 UNet2DS.....	126
References .....	127

## List of Tables

Table 1: List of one-sided distributions used in model fitting.....	17
Table 2: List of values used for the known parameters in simulation test. ....	21
Table 3: List of chosen mean and standard deviations for prior distributions. ....	22
Table 4: Description of data used from the Allen Brain Observatory. All data are from the primary visual cortex.....	45
Table 5: Summary of performances on all datasets.....	66
Table 6: STNeuroNet performance on all data when trained on different datasets.....	73
Table 7: GCL soma detection performance scores for the IU dataset.....	100
Table 8: Effect of intensity normalization and test-time-augmentation on detection performance for IU's dataset.....	101
Table 9: GCL soma detection performance scores on the FDA dataset. ....	105
Table 10: Effect of intensity normalization and test-time-augmentation on detection performance for FDA's dataset.....	106
Table 11: Generalizability test between groups of subjects from the FDA dataset. ....	108
Table 12: Generalizability of trained method across different AO-OCT imaging systems with different scan characteristics. ....	112

## List of Figures

Figure 1: Closely timed AP induced fluorescence transients accumulate, making the detection of individual spikes a challenge. ....	9
Figure 2: Overview of the single spike waveform extraction and curve fitting for characterizing the prior probability model. ....	16
Figure 3: The smallest detectable ISI ( $ISI_{\min}$ ) determined the resolution limit.....	26
Figure 4: Two spike detection results from the experimental dataset.....	29
Figure 5: Lower bounds on ISI estimation at $f_s = 500$ Hz for GCaMP6s and GCaMP6f. ...	30
Figure 6: ML estimators nearly achieved the information-theoretic bounds. ....	34
Figure 7: Histograms of experimentally calculated ISI values in the $\alpha = \beta$ known case and combined SNR = 20. ....	35
Figure 8: MAP estimators nearly achieve the information-theoretic bounds. ....	36
Figure 9: Overlapping neurons complicate active neuron segmentation. ....	42
Figure 10: Schematic for the proposed spatiotemporal deep learning-based segmentation of active neurons in two-photon calcium videos. ....	49
Figure 11: The optimal thresholds in the post-processing step of our algorithm are determined through leave-one-out cross-validation. ....	53
Figure 12: Representative examples of active neuron labeling errors in the ABO dataset. ....	61
Figure 13: STNeuroNet accurately identified active neurons from the ABO dataset. ....	63
Figure 14: STNeuroNet outperformed other methods on the ABO dataset. ....	65
Figure 15: Inter-human agreement test for ABO neuron segmentation.....	67
Figure 16: Trained STNeuroNet performed equally well on data from a different cortical layer and outperformed other methods. ....	70
Figure 17: STNeuroNet achieved best performance in the Neurofinder challenge, which contained suboptimal markings. ....	71

Figure 18: Overview of the weakly supervised deep learning method for GC segmentation.....	86
Figure 19: Extraction of the ganglion cell layer (GCL) from AO-OCT volumes.....	91
Figure 20: Network architecture .....	94
Figure 21: Unsupervised segmentation of GCL somas using the CNN’s learned features. ....	96
Figure 22: Our method achieved expert-level performance across different retinal locations on the IU dataset.....	102
Figure 23: Comparison between automatic and manual GCL soma segmentations. ....	103
Figure 24: Illustrative results on the IU dataset.....	104
Figure 25: Results on FDA’s healthy and glaucoma subjects. ....	107
Figure 26: Illustrative results on FDA’s dataset.....	109
Figure 27: Structural and functional characteristics of glaucomatous eyes compared to controls. ....	111

## Acknowledgements

I would like to acknowledge the support and encouragement I received during my PhD research. I am sincerely grateful to my advisor, Dr. Sina Farsiu, for his tremendous support, critical and open-minded scientific inputs, and excellent mentorship. My earnest thanks to Dr. Yiyang Gong for his support, mentorship, and constructive criticism in my dissertation. I thank my dissertation committee members Dr. Joseph A. Izatt, Dr. Warren M. Grill, and Dr. Junjie Yao for their scientific and critical contributions. I would like to thank my colleagues and friends at the Vision and Image Processing Lab, Dr. David Cunefare, Jessica Loo, Ali Hasan, Daniel Park, Ziyun Yang, Kevin Choy, and Leon Kwark, and members of the Neurotoolbox Lab, Emily Redington, Connor Beck, Dr. Diming Zhang, Dr. Depeng Wang, Dr. Yijun Bao, Jaebin Kim, Yuqi Tian, Kimberly Lennox, and Dr. Joshua Khani for their support, motivation, and scientific inputs which significantly helped my projects. I appreciate our collaborators from Indiana University, Dr. Donald T. Miller and Dr. Kazuhiro Kurokawa, and FDA, Dr. Daniel X. Hammer, Dr. Zhuolin Liu, and Dr. Furu Zhang for their scientific inputs and providing access to their data. I acknowledge the financial support of the Pre-doctoral NIH Fellowship in the Medical Imaging Training Program (T32-EB002040), the NSF BRAIN Initiative (NCS-FO 1533598), and the 2020 Unrestricted Research to Prevent Blindness Grant award in my projects.

Finally, I acknowledge the people who mean a lot to me. I am incredibly grateful to my parents and sisters for their unconditional support and for inspiring me to pursue my dreams. To my husband, thank you for your emotional support, encouragement,

patience and understanding throughout the process, this journey would have not been possible without you.

# 1. Introduction

## *1.1 Motivation*

Understanding how neuronal activities drive cognition and behavior and how functionality is impaired during the diseased states are among the greatest challenges in neuroscience. The underlying biological basis of many normal and impaired functional processes is still unknown due to the complexity of the central nervous system. To add new knowledge to the field, minimally invasive recording techniques with single cell resolution are critical. Optical recording techniques allow high resolution imaging of neurons in the central nervous system, e.g. the brain and retina.

Currently neuroscientists are using the combination of optical recording techniques and genetically encoded calcium indicators (GECIs) to image neuronal activities with high spatiotemporal resolution in animal brains. Beyond recording neuronal activity, precise mapping of different neuron populations' contribution to circuit-level brain activity and behavior can be achieved by utilizing optogenetic tools [1]. Most published optogenetic experiments in behaving animals, especially those with simultaneous optical neuronal readout, don't receive real-time feedback from the recordings to guide the stimulation hardware [1-3]. Full realization of an all optical closed loop optogenetic experiment has been impeded due to the lack of fast and accurate computational frameworks that can drive the optical components of the system. Thus, to fully understand how stimuli are transmitted among neurons by real-time optical manipulations, we need image and signal processing algorithms with high accuracy, precision, and speed in parsing neuronal activities. Other types of experiments may also

benefit from fast data processing frameworks. For example, single day experimental paradigms that are a series of behavioral sessions interleaved with rest sessions will have the opportunity to plan or modify the following behavioral session based on the results of the current session. This opportunity can be provided only if the processing frameworks yield the required numerical results within the rest session, which typically last for minutes.

Optical imaging systems can monitor neuronal activities from hundreds to thousands of neurons. Recording this amount of information in behavioral experiments that typically last about tens of minutes, result in gigabytes of data. To extract neuronal activity traces, regions of interest (ROIs) corresponding to neuron somas need to be identified. ROI segmentation from this amount of data is challenging, especially if it is done manually. Additionally, optical recording of neuronal activity with GECIs suffers from limited temporal resolution due to the slow dynamics of the indicator. Thus, accurate spike detection algorithms are needed to infer action potentials from these traces.

Current processing methods for calcium imaging data can be categorized into three groups: 1) Methods that infer spike trains from neuronal traces assuming known spatial maps of neurons (e.g., determined manually) [4-12], 2) methods that segment neurons by processing a summary image that aggregates temporal information (e.g., the mean image) [13], and 3) methods that jointly segment neurons and infer spike trains [14-20].

In terms of spike time estimation, closely timed action potential evoked fluorescence transients accumulate, making the detection of individual spikes a challenge. Over the past years, several groups have tackled the problem of spike train extraction or firing rate inference from the observed fluorescence signals. However, there is a lack of theoretical analysis on the resolution limit of resolving two closely timed fluorescence traces. Such analysis could aid in future spike detection methods and serve as a reference for evaluating the capability of computational methods in detecting closely timed action potentials.

From the three groups of processing methods, methods from the third group are superior because they exploit the full spatiotemporal information in the data. Based on the high temporal correlation between pixels belonging to an individual neuron, Mukamel et al. [14] proposed an independent component analysis (ICA) based method to segment neurons. However, this method will fail in large field-of-view recordings where spatially distant neurons can have high temporal correlations. Pnevmatikakis et al. [16] proposed constrained matrix factorization (CNMF) to simultaneously segment neurons and deconvolve their temporal traces. Compared to ICA, this method is more robust in low signal-to-noise ratio (SNR) cases [16], but is highly sensitive to initialization. Both methods have the disadvantages of requiring the entire data to be available and take hours to converge. These methods also require determining the number of neurons ( $N$ ) prior to the analysis, which is unknown in general. Underestimating  $N$  will result in missing neurons, and overestimation of  $N$  will falsely detect non-neural structures as neurons.

Therefore, these methods require human evaluation or automated post-processing pipelines to validate the end results. These drawbacks make such methods unsuitable for experiments that require fast data processing frameworks such as conditional stimulation based on real-time readout of neural activity.

Taking advantage of online dictionary learning, Giovannucci et al. introduced OnACID [19], which processes the video as it becomes available and does not require  $N$  to be determined beforehand. More recently, Giovannucci et al. [18] have improved the scalability of CNMF and extended OnACID with new initialization methods and a convolutional neural network (CNN), referred to as CaImAn Batch and CaImAn Online, respectively. These improvements have increased the processing speed, however, as we will show in chapter 3, there is a gap between human-level and the methods' performances.

Shifting the focus from the brain to the retina, successful diagnosis, prognosis, and treatment of many retinal neurodegenerative diseases, including glaucoma, Parkinson's disease, and Alzheimer's disease depend on the visualization of microscopic retinal structures. The incorporation of adaptive optics (AO) with the high-resolution optical coherence tomography (OCT) allows volumetric visualization of the retina at cellular resolution. In glaucoma, the second leading cause of blindness worldwide [21, 22], the loss of retinal ganglion cells (GCs) is directly associated with the loss of the visual field [23]. Recently, GC layer somas have been visualized with AO-OCT systems [24, 25]. Currently, manual marking of AO-OCT volumes is the standard approach for quantifying the GC

population, which is subjective, time consuming, and not practical for large-scale studies and clinical use. Thus, there is a critical need for an automated, high throughput GC detection and sizing algorithm from AO-OCT volumes.

In this dissertation, in addition to presenting statistical and information-theoretic analysis tools for neuronal fluorescence traces, automatic methods for rapid and high throughput segmentation of neuronal images were developed. Our computational frameworks will aid in new discoveries in basic neuroscience, such as adaptation, plasticity, and neural state changes, or clinically relevant illnesses. More specifically, due to the computational speed of our proposed frameworks, it will facilitate all optical, closed-loop optogenetic experiments where rapid parsing of neuronal activity is needed. Additionally, the developed frameworks will aid in repeatable measures of retinal GCs and thus, improve retinal neurodegenerative care and potentially facilitate treatment strategies

## ***1.2 Innovation***

The developed frameworks in this dissertation are innovative in approach. We have built upon modern information theory and machine learning methods to accurately convert optical recordings of the brain and the retina into desired quantitative features. In the context of *in vivo* two-photon microscopy studies, the frameworks are innovative in that they rapidly parse neuronal activity based on learned features with little dependence on recent information, allowing for closed-loop, simultaneous optical recording and manipulation experiments. By taking advantage of supervised learning techniques, we

developed generalizable calcium imaging data analysis frameworks without the need for excessive parameter tuning per recorded data.

On another front, the frameworks developed in this dissertation are innovative with respect to combining the advanced retinal imaging modality of AO-OCT with novel deep learning techniques to accurately quantify morphological properties of retinal neurons. Aside from novelty in application, the proposed deep learning-based methods are innovative in network architecture and the CNN for GC segmentation is innovative in that it segments individual GC somas using weakly-supervised training, requiring significantly less human input for the training process. Our GC soma segmentation framework is the first fully automatic approach for volumetric processing of AO-OCT images, leading to a shift in clinical and research practice in the future.

## 2. Fundamental Limits in Resolving Neural Spikes in Two-photon Calcium Imaging Recordings

Although optical imaging of neurons using fluorescent genetically encoded calcium sensors has enabled large-scale *in vivo* experiments, the sensors' slow dynamics often blur closely timed action potentials into indistinguishable transients. While several previous approaches have been proposed to estimate the timing of individual spikes, they have overlooked the important problem of estimating inter-spike-interval (ISI) for overlapping transients. The purpose of this chapter was to quantify the resolution limits of detecting closely timed fluorescent transients using statistical and information theoretic tools. The methods developed in this chapter were published in *IEEE Transactions of Biomedical Engineering* under the title "Information-theoretic approach and fundamental limits of resolving two closely timed neuronal spikes in mouse brain calcium imaging" [26], and presented at the *Biomedical Engineering Society annual meeting* under the title "Statistical analysis and performance limits of inter-spike interval estimation in calcium imaging". Thus, the content of this chapter (text, figures, tables, and equations) were mainly reproduced from these publications.

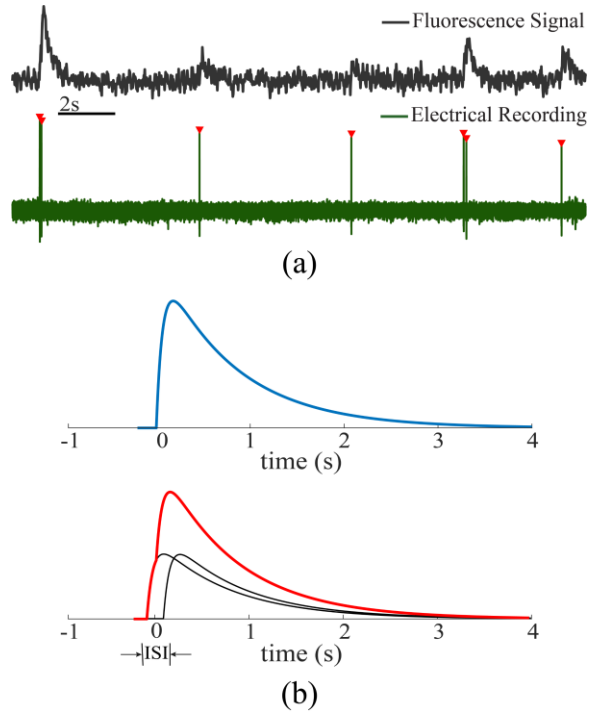
### 2.1 Introduction

Recent advances in optical microscopy and genetically encoded calcium indicators (GECIs) have increased the use of these tools in large scale *in vivo* recording of neuronal populations [2, 27, 28]. Accurate extraction of neuronal activities from the optical recordings is expected to give insight into how neuronal circuitry process information.

Therefore, to fully understand how stimuli are processed and transmitted among neurons, spike extraction approaches with high accuracy and precision are needed. However, during periods of rapid activity, closely timed AP induced fluorescence transients accumulate, making the detection and separation of individual spikes a challenge (Figure 1(a)).

Over the past years, several groups have tackled the problem of firing rate inference or spike train extraction from the observed fluorescence signals. Methods that estimate firing rate or spiking probabilities include fast nonnegative deconvolution [29], supervised learning with probabilistic models [30], and the Markov Chain Monte-Carlo methods [31]. Methods that estimate spike trains include nonnegative deconvolution [16], sparsity-based reconstruction [32-35], template matching [8, 36, 37], finite rate of innovation [6, 38], and Bayesian methods [4].

Several of these past studies behave well in reconstructing neuronal bursting activities [30, 33-35, 38]. However, few of these studies have conducted theoretical performance limit analysis. Such analysis can aid in resolving experimental design issues for optimal spike detectability, such as sensor kinetics, recording speed, and photon counts [37]. Among the above-mentioned studies, only [37] and [38] have compared their single spike time estimation method with the optimal performance of any unbiased estimator through the Chapman-Robbins and Cramer-Rao lower bounds (CRB), respectively.



**Figure 1: Closely timed AP induced fluorescence transients accumulate, making the detection of individual spikes a challenge. Simultaneous optical and electrical recording from a neuron. Red markers show spike times. (b) Illustration of the null hypothesis with one AP-induced signal (top), and the alternative hypothesis of two closely timed fluorescence signals with sub-second ISI (bottom). Figure taken from [26]. © [2018] IEEE**

Here we extend the application of previous studies [37, 38] which considered only isolated spikes, by investigating the case of temporally overlapping waveforms. In parallel to other computational spike extraction methods, we quantify 1) resolution from the statistical point-of-view and 2) the theoretical bound on the precision of estimating the ISI. Our work is based on the statistical and information theoretic tools developed in the past two decades for estimating the fundamental resolution of optical systems [39-43]. As the symmetric point spread function (PSF) considered in the numerical results of these

optically oriented papers does not match our problem, we extend this framework for the ISI estimation and study possible consequences of asymmetric waveforms.

Experiments based on simulated and real data across different SNR levels and recording speeds showed that our algorithms can accurately distinguish two fluorescence signals with ISI on the order of tens of milliseconds, shorter than the waveform’s rise time. Our study showed that the statistically optimal ISI estimators closely approached the CRBs. Such analysis aids not only in future spike detection methods, but also in future experimental design when choosing sensors of neuronal activity.

## 2.2 Methods

### 2.2.1 Action Potential Evoked Fluorescence Signal Model

In response to an AP, the intracellular calcium concentration rises rapidly, which is followed by a slow decay to its baseline value [44]. As validated by experiments in [45], we assume that the measured fluorescence signal is linearly related to the intracellular calcium concentration. Given samples at time points  $t_k (k = 1, 2 \dots, K)$ , the mean fluorescence signal generated in response to a single AP at time  $t = 0$  with normalized peak amplitude  $\theta_0 = A$  is expressed as [36]

$$s_0(t_k; \theta_0) = AF_0 h(t_k) + F_0, \quad (1)$$

where the change in the fluorescence signal is modeled as

$$h(t_k) = a (1 - \exp(-t_k/\tau_{on})) \exp(-t_k/\tau_d) u(t_k). \quad (2)$$

In Eq. 1,  $F_0$  is the baseline photon rate due to the neuron’s resting state fluorescence, auto-fluorescence from cellular structures, and fluorescence from the extracellular space. In Eq.

2,  $u(t_k)$  is the unit step function,  $a$  is a normalization factor, and  $\tau_{on}$  and  $\tau_d$  are known rise and decay time constants, respectively. Assuming the optical technique used for measurement has negligible read out noise, the recordings are photon shot noise limited. Therefore, the  $K$ -element measurement vector  $\mathbf{y}$  is distributed according to Poisson statistics with a time-varying mean  $s_0(t_k; \theta_0)$ .

## 2.2.2 Statistical Analysis of Resolution

We test the hypothesis of whether one or two spikes are present at an observation window of length  $K$  points, as illustrated in Figure 1(b). The null hypothesis  $H_0$  denotes the case where there is one spike present as described in the previous section. The alternative hypothesis  $H_1$  refers to the case where we have two spikes with  $ISI \neq 0$ . The peak amplitude of the signal generated by the two spikes in  $H_1$  should be comparable to the peak amplitude of a single spike under the  $H_0$  hypothesis. Thus, in the case where the neuron spikes twice at times  $d_1$  and  $-d_2$  ( $ISI = d_1 + d_2 = d$ ) with normalized peak amplitudes  $\alpha$  and  $\beta$  (where  $A = \alpha + \beta$ ), we define the accumulated mean signal as

$$s_1(t_k; \theta_1) = \alpha F_0 h(t_k - d_1) + \beta F_0 h(t_k + d_2) + F_0, \quad (3)$$

where  $\theta_1 = [d_1, d_2, \alpha, \beta]$  is the parameter vector defining the signal.

The probability distribution of the set of photon measurements  $\mathbf{y}$  under  $H_j$  ( $j \in [0, 1]$ ) is then modelled as

$$p(\mathbf{y}|H_j) = \prod_{k=1}^K \text{Poisson}(s_j(t_k; \theta_j)) = \prod_{k=1}^K \exp(-s_j(t_k; \theta_j)) \frac{s_j^{y(t_k)}(t_k; \theta_j)}{y(t_k)!}. \quad (4)$$

The Log-likelihood Ratio Test (LRT) can be used to choose between the two hypotheses [46] for a given set of measurements:

$$L(\mathbf{y}) = \ln\left(\frac{p(\mathbf{y}|H_1)}{p(\mathbf{y}|H_0)}\right) = \sum_{k=1}^K y(t_k) \ln\left(\frac{s_1(t_k; \boldsymbol{\theta}_1)}{s_0(t_k; \boldsymbol{\theta}_0)}\right) - \sum_{k=1}^K (s_1(t_k; \boldsymbol{\theta}_1) - s_0(t_k; \boldsymbol{\theta}_0)) \quad (5)$$

For any given dataset,  $H_1$  is selected as the more likely hypothesis if the log-likelihood ratio exceeds a predefined threshold. The choice of threshold depends on the desirable value for the probability of detection,  $P_D$ , or the tolerable value for the probability of false positive,  $P_F$  [46].

Parameters ( $\tau_{on}$ ,  $\tau_d$ , and  $F_0$ ) for a specific calcium probe in a biological system can be systematically characterized and thus are assumed to be known quantities. However, the model parameters ( $[\theta_0, \theta_1]$ ) in the above LRT are unknown in general. To address this composite hypothesis problem, we use the Generalized Likelihood Ratio Test (GLRT) to simultaneously assess the existence of two spikes and estimate the ISI between them. GLRT uses the maximum likelihood (ML) estimates of the unknown parameters to form the Neyman-Pearson detector [46]. The ML estimates of the unknown parameters in  $\theta_j$  are found by maximizing the log-likelihood function of the data under  $H_j$  ( $j \in [0, 1]$ ).

$$\hat{\theta}_j = \operatorname{argmax}_{\theta_j} [\ln(p(\mathbf{y}|H_j))] = \operatorname{argmax}_{\theta_j} \sum_{k=1}^K \left[ y(t_k) \ln(s_j(t_k; \boldsymbol{\theta}_j)) - s_j(t_k; \boldsymbol{\theta}_j) \right], \quad (6)$$

where we have kept only the parameter dependent parts. We numerically solve the above nonlinear maximization problem using MATLAB's optimization toolbox. Note that without loss of generality, we have set the single spike model characterized in Eq. 1 to start at  $t = 0$ . In the case of aiming to detect the timing of single spikes, modifying the

signal model in Eq. 2 to include the unknown time shift and using this model in the maximum likelihood equation will solve the problem. Before addressing the general case of detecting spikes with fully unknown parameters, we consider the more intuitive case of detecting spikes with known amplitudes, as in [41].

### 2.2.2.1 Spikes with known amplitudes

The hypotheses for differentiating the case of one large spike starting at the test origin (defined as time zero), versus the case of two smaller amplitude spikes located around the test origin are expressed as

$$H_0: s_0(t_k) = (\alpha + \beta)F_0h(t_k) + F_0, \quad (7)$$

$$H_1: s_1(t_k; \boldsymbol{\theta}_1) = \alpha F_0 h(t_k - d_1) + \beta F_0 h(t_k + d_2) + F_0, \quad (8)$$

$$\boldsymbol{\theta}_1 = [d_1, d_2].$$

Note that while the amplitudes  $(\alpha, \beta)$  are assumed to be known, their values can be equal or different. Also, the spikes in the  $H_1$  case can be symmetrically ( $d_1 = d_2 = d/2$ ) or asymmetrically ( $d_1 \neq d_2$ ) distributed around the test origin.

The minimum detectable distance between two spikes that can be distinguished from a single spike is modified by how the time origin of the test is defined. Conceptually, the most challenging problem set up has high temporal overlap between  $s_1(t_k; \boldsymbol{\theta}_1)$  and  $s_0(t_k)$ . Numerically, such a set up can be attained by finding the maximum point of cross-correlation between  $s_1(t_k; \boldsymbol{\theta}_1)$  and  $s_0(t_k)$  [41]. This setup, using the Taylor expansion, then defines the test time origin  $\tau$  as

$$|\tau| = \frac{\alpha d_1 - \beta d_2}{\alpha + \beta}. \quad (9)$$

Therefore, fixing the location of the test origin to  $\tau = 0$  leads to  $\alpha d_1 = \beta d_2$ , or equivalently,

$$d_1 = \frac{\beta}{\alpha + \beta} d \text{ and } d_2 = \frac{\alpha}{\alpha + \beta} d. \quad (10)$$

This suggests that, for  $\alpha = \beta$ , the “best” (i.e. most challenging) location to carry out the hypothesis test is in the middle of the two transients and for  $\alpha \neq \beta$ , the point should be closer to the larger signal. The condition of whether the spikes are symmetrically or asymmetrically located around the test origin is studied to investigate the effect of defining the test origin, or equivalently the  $H_0$  hypothesis, in quantifying the resolution limit.

### 2.2.2.2 Spikes with random amplitudes

Calcium ion influx through calcium channels and calcium binding to the sensor are stochastic processes that can lead to variations in the calcium signal response and thus, the fluorescence signal. Therefore, the signal peak value of a single spike can change from time to time and even drastically from one neuron to another. To encompass these variabilities, we consider the more general case of differentiating spikes with unknown intensities, by treating the peak amplitudes as random variables. The  $H_0$  hypothesis is described by Eq. 1, and the  $H_1$  hypothesis under the condition in Eq. 10 is expressed as

$$H_1: s_1(t_k; \boldsymbol{\theta}_1) = \alpha F_0 h(t_k - d_1) + \beta F_0 h(t_k + d_2) + F_0, \quad (11)$$

$$\boldsymbol{\theta}_1 = [d, \alpha, \beta].$$

This is a generalization of the previous work, in which the amplitudes of unknown signals were assumed to be deterministic [41]. Since a Bayesian hypothesis testing approach to

combine observation data and *a priori* information about the peak amplitude distribution involves integrations that are not analytically solvable, we used the GLRT based conditional ML estimation technique [47]. We incorporate the prior information in quantifying the performance of the detector through computing the expected value of  $P_D$  (and  $P_F$ ) over  $p(\alpha, \beta)$ , the joint probability distribution of the amplitudes [47].

The prior probability distribution of the single spike amplitude has not been previously investigated. Therefore, we set to find the best probability model from a set of measurements.

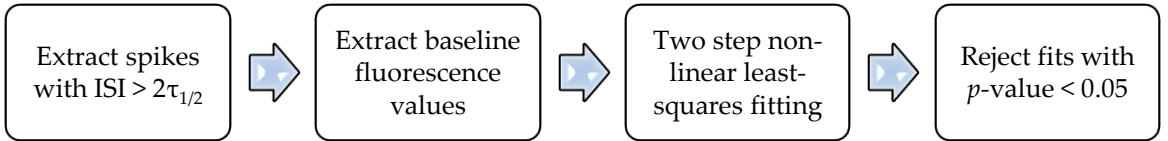
### **2.2.3 Extraction of Single and Double AP Evoked Fluorescence Transients**

We used the publicly available experimental dataset, provided by the Svoboda lab [45] as reference. The dataset contains simultaneous optical imaging and loose-seal cell-attached recording of nine GCaMP6s and eleven GCaMP6f (types of GECIs) expressing neurons. We extract single AP induced transients to find the best probability model for peak fluorescence response. The outline of the processing steps is highlighted in Figure 2. First, we identify single fluorescence transients using the electrophysiological data and extract them from the optical recordings. To ensure accurate estimation of single AP evoked fluorescence peak values, we discard spikes with ISI values less than twice the fluorescence half decay time constant ( $2\tau_{1/2}$ , approximately 1s for GCaMP6s and 0.3 s for GCaMP6f [45]). We also discard cases with high neuropil contamination. Second, the background signal  $F_0$  for each spike is calculated by averaging the baseline near the onset

time in periods with no neuronal activity. Next, we use a two-step nonlinear least square procedure to fit a double exponential model as in Eq. 2 to the extracted spike transients. The least square curve fitting method finds the best fit of the model to the data,  $y_i$ , by minimizing

$$\chi^2 = \sum_{i=1}^n (y_i - f_i)^2 / \sigma_i^2, \quad (12)$$

where  $f$  is the set of estimated values and  $\sigma_i^2$  is the standard deviation (std) of each data point [48]. Since the data has Poisson statistics, in the first step, we use the data itself as an estimate of  $\sigma_i^2$ . We repeat the fitting for a second time to reduce the overemphasis of data points with lower variance [49]. In this step, we use the fitted values,  $f_i$ , as the estimates of  $\sigma_i^2$ . Lastly, we evaluate the goodness of fit by calculating the  $p$ -value associated with the final  $\chi^2$  value. Signals that have a poor fit ( $p$ -value  $< 0.05$ ) to the model, are discarded from further analysis.



**Figure 2: Overview of the single spike waveform extraction and curve fitting for characterizing the prior probability model.**

We use the fitted results to obtain the distribution of normalized peak values for each neuron. We test fifteen different one-sided distributions listed in Table 1 on all neurons separately. We determine the best distribution model among all neurons using a two-step procedure. First, for each neuron, we calculate the ML estimates of each model’s parameters. We then use Pearson’s  $\chi^2$  goodness of fit test for each fitted model. Models

that result in  $p$ -values  $< 0.05$  are discarded from the set of possible probability models. In the second step, we choose the best probability model among the remaining models using the Akaike Information Criterion (AIC), defined as [50]

$$\text{AIC} = -2\ln(f(x|\hat{\theta})) + 2k, \quad (13)$$

where  $\hat{\theta}$  is the ML estimates of the model's  $k$  parameters based on the observations,  $x$ . For a single dataset, the model resulting in the smallest AIC score is the best model that represents the data [50]. We select the probability model with the lowest sum of AIC score across all neurons as the model that best describes the dataset (among the considered models).

**Table 1: List of one-sided distributions used in model fitting. Table taken from [26]. © [2018] IEEE**

#	<i>Distribution Name</i>	#	<i>Distribution Name</i>
1	Rayleigh	9	Log-Normal
2	Birnbaum-Saunders	10	Nakagami
3	Extreme Value	11	Normal
4	Gamma	12	Rician
5	Half Normal	13	Weibull
6	Inverse Gaussian	14	Burr
7	Logistic	15	T Location Scale
8	Log-Logistic		

We also extract visually indistinguishable double spike cases to demonstrate the detection performance of our framework on experimentally obtained data. Double spike signals are defined as cases with two closely-timed spikes without any other spike occurring within  $2\tau_{1/2}$  time interval around them. Further, we discarded cases in which the two spikes were visually distinguishable. We centered the two spike signals such that

time  $t = 0$  is in the middle of the two waveforms. Examples of one and two spike signals are illustrated in Figure 1(a).

## 2.2.4 The Cramer-Rao Lower Bound

In this section, we utilize the Cramer-Rao based lower bounds as reference to study the limits of attainable precision in the estimation of the AP evoked fluorescence transient peak amplitudes and ISI under  $H_1$  hypothesis. The covariance matrix  $\mathbf{C}$  of any unbiased estimator of the  $p$ -parameter vector  $\boldsymbol{\theta}_1$  is a  $p \times p$  matrix that satisfies [51]

$$\mathbf{C}_{\hat{\boldsymbol{\theta}}_1} \geq \mathbf{I}_F^{-1}, \quad (14)$$

where  $\mathbf{I}_F$  is the  $p \times p$  Fisher information matrix. The elements of  $\mathbf{I}_F$  for data with Poisson statistics are calculated as

$$\mathbf{I}_{F_{ij}} = -\mathbb{E} \left\{ \frac{\partial^2 p(\mathbf{y}; \boldsymbol{\theta}_1)}{\partial \boldsymbol{\theta}_{1i} \partial \boldsymbol{\theta}_{1j}} \right\} = \sum_{k=1}^K \frac{1}{s_1(t_k; \boldsymbol{\theta}_1)} \frac{\partial s_1(t_k; \boldsymbol{\theta}_1)}{\partial \boldsymbol{\theta}_{1i}} \frac{\partial s_1(t_k; \boldsymbol{\theta}_1)}{\partial \boldsymbol{\theta}_{1j}}. \quad (15)$$

In the following sections, the CRBs for estimating ISI,  $\alpha$ , and  $\beta$  are derived for the problems described in sections 2.2.2.1 Spikes with known amplitudes and 2.2.2.2 Spikes with random amplitudes.

### 2.2.4.1 CRB for known amplitude signals

For this case, under the assumption of Eq. 10, the only unknown parameter is  $d$ .

Therefore, the Fisher information matrix reduces to a scalar calculated as

$$\mathbf{I}_F = -\mathbb{E} \left\{ \frac{\partial^2 p(\mathbf{y}; d)}{\partial d^2} \right\} = \sum_{k=1}^K \frac{1}{s_1(t_k; d)} \left( \frac{\partial s_1(t_k; d)}{\partial d} \right)^2. \quad (16)$$

Thus, the lower bound for the unbiased estimation of  $d$  is  $\text{var}(\hat{d}) \geq I_F^{-1}$ . We refer to Appendix A for the full derivation of the above quantity.

#### 2.2.4.2 Hybrid CRB for random amplitude signals

To address the more challenging case of random spike amplitudes combined with unknown deterministic ISI, we estimate the unknown parameters through a joint ML and maximum a posteriori (MAP) estimator. This optimization problem involves the simultaneous ML estimation of ISI (or  $d$ ) and MAP estimation of the normalized peak amplitudes [47] ( $\alpha$  and  $\beta$ ):

$$\hat{\theta}_1 = \begin{bmatrix} \hat{d}_{ML} \\ \hat{\alpha}_{MAP} \\ \hat{\beta}_{MAP} \end{bmatrix} = \underset{d, \alpha, \beta}{\text{argmax}} [\ln p_{y|\theta_1}(y|\theta_1) + \ln p_{\alpha, \beta|d}(\alpha, \beta|d)], \quad (17)$$

where  $p_{\alpha, \beta|d}(\alpha, \beta|d)$  is the conditional joint prior distribution of the amplitudes. For this hybrid problem, we utilize the more general Hybrid CRB (HCRB) [47] method, which is defined as

$$\text{HCRB} \geq \mathbf{I}_H^{-1}, \quad (18)$$

$\mathbf{I}_H$  is called the Hybrid information matrix, which defines the lower limit on the mean square error (MSE) of any estimator. It is a  $3 \times 3$  matrix for the problem in section 2.2.2.2

Spikes with random amplitudes ( $\theta_1 = [d, \alpha, \beta]$ ), expressed as the sum

$$\mathbf{I}_H = \mathbf{I}_D + \mathbf{I}_P, \quad (19)$$

where

$$\mathbf{I}_D(d) = E_{\theta_r|d}[\mathbf{I}_F(d, \theta_r)], \quad (20)$$

and

$$\mathbf{I}_{P_{ij}} = -\mathbb{E}_{\theta_r|d} \left[ \frac{\partial^2 \ln p(\theta_r|d)}{\partial \theta_{r_i} \partial \theta_{r_j}} \right], \quad (21)$$

$$\theta_r = [\alpha, \beta].$$

The elements of the Fisher information matrix are calculated according to Eq. 15, in which the derivatives of the mean signal  $s_i(t_k; \theta_i)$  relative to the amplitudes are derived in Appendix A. To attain  $\mathbf{I}_D$ , we calculate the expectation of  $\mathbf{I}_F$  with respect to  $\alpha$  and  $\beta$ . Note that the amplitudes of the two spikes are independent and identically distributed (i.i.d) random variables and independent from  $d$ , i.e.  $p(\alpha, \beta|d) = p(\alpha)p(\beta)$ . This integral is numerically solved using MATLAB.  $\mathbf{I}_P$  on the other hand, can be attained analytically, which is derived in section 2.3.3 Prior Knowledge about Signal Amplitudes Yields Theoretically Equal ISI Estimation Performance to the Known Case based on the best model match for the prior distribution of  $\alpha$  and  $\beta$ .

## 2.3 Results

Numerical analysis of the minimum detectable ISI and the CRBs through biologically plausible simulations and using experimental data are presented in this section. Our simulations are parameterized based on the experimental results in [5, 45] for two different calcium sensors: GCaMP6s and GCaMP6f. Since multiple existing scanning techniques have different imaging speeds, we consider multiple frame rates ( $f_s$ ) in our simulations as well. Acousto-optical deflector (AOD) based two-photon microscopes have allowed high speed imaging of neuronal activities up to 500 Hz [5], enabling millisecond precision spike time estimations. Resonant scanning methods are more widely used,

achieving 30 Hz for a 512×512 pixels field-of-view, or 60 Hz for a smaller area such that the laser dwell time per neuron is approximately kept the same. Without loss of generality, we consider the case in which the dwell time per neuron is constant across different recording speeds for the comparison between their resolution limits and theoretical lower bounds. Table 2 lists the values of parameters used in the simulations. We determine the dwell time by considering a 15 μm diameter neuron imaged by systems with 1 μm pixel size.

**Table 2: List of values used for the known parameters in simulation test. Table taken from [26]. © [2018] IEEE**

<i>Parameter</i>	<i>GCaMP6s</i>	<i>GCaMP6f</i>
$\tau_{on}$	72 ms [45]	18 ms [45]
$\tau_d$	793.5 ms [45]	204.9 ms [45]
$f_s$	500 Hz (AOD), 60 Hz and 30 Hz (Resonant)	500 Hz (AOD), 60 Hz and 30 Hz (Resonant)
Dwell time	25 μs	25 μs

### 2.3.1 The Gamma Distribution Characterizes the Peak Amplitude

The data extraction pipeline described in section 2.2.3 Extraction of Single and Double AP Evoked Fluorescence Transients resulted in  $n = 44, 10, 13, 51, 48, 30, 61, 10,$  and 13 waveforms per GCaMP6s labeled neurons and  $n = 100, 63, 88, 39, 14, 60, 99, 283, 54, 38,$  and 93 waveforms per GCaMP6f labeled neurons. The  $\chi^2$  test eliminated distribution numbers 1, 10, 11, 12, 13, 14, and 15 for GCaMP6s neurons and 1, 2, 3, 5, 6, 8, 9, and 15 for GCaMP6f neurons from Table 1. Among remaining models, the Gamma distribution

resulted in the minimum sum of AIC score for both calcium sensors. The Gamma probability distribution with parameters  $k$  and  $c$  is defined as [52]

$$f(x; k, c) = \frac{c^{-k} x^{k-1} \exp\left(-\frac{x}{c}\right)}{\Gamma(k)}, \quad x > 0 \quad (22)$$

where  $\Gamma(k)$  is the gamma function with argument  $k$ . The mean and variance of this distribution are  $kc$  and  $kc^2$ , respectively [52].

Similar to previous calcium imaging studies [5], we define signal-to-noise ratio (SNR) as

$$\text{SNR} = \left(\frac{\Delta F}{F_0}\right) \sqrt{F_0}, \quad (23)$$

where  $\Delta F$  is the change in fluorescence of one AP evoked calcium transient at its peak amplitude, equal to  $AF_0$  in Eq. 1. Noting that the mean and variance of the Gamma distribution are dependent, we carry out simulations with different levels of SNR by fixing  $k$  and  $c$  (thus fixing the mean and variance) while changing the baseline photon rate  $F_0$ . Based on the mean and standard deviation of  $\Delta F/F$  values from all neurons in each dataset, we selected the mean and standard deviation of both sensors'  $\Delta F/F$  prior distributions as listed in Table 3. To be consistent in simulations between the known and unknown amplitude cases, we carried out the simulations related to section 2.2.2.1 Spikes with known amplitudes with  $\alpha + \beta = 0.46$  for GCaMP6s and  $\alpha + \beta = 0.38$  for GCaMP6f.

**Table 3: List of chosen mean and standard deviations for prior distributions.**

<i>Parameter</i>	<i>GCaMP6s</i>	<i>GCaMP6f</i>
Mean	0.23	0.19
Std	0.03	0.06

## 2.3.2 The Detector Distinguishes Two Fluorescence Transients with ISI on the Order of Tens of Milliseconds

### 2.3.2.1 Performance characterized through data simulation

Due to the asymmetry of the transients, ISI values greater than  $t_{rise}$  (the time when the fluorescence transient  $h(t)$  reaches its maximum) result in visually distinguishable transients. Therefore, we are interested in the range of values  $ISI < t_{rise}$ . For the bi-exponential model described in Eq. 2 and according to GCaMP6s parameters in Table 2:

$$t_{rise} = \tau_{on} \ln \left( 1 + \frac{\tau_d}{\tau_{on}} \right) \cong 0.2 \text{ s.} \quad (24)$$

Numerical evaluation of the smallest detectable ISI depends on the selection of  $P_D$  and  $P_F$ . We set the number of false positives to be equal to the number of misses, relating  $P_F$  and  $P_D$  through

$$P_F = \frac{p(H_1)}{1 - p(H_1)} (1 - P_D), \quad (25)$$

where  $p(H_1)$  is the probability of the  $H_1$  hypothesis. Assuming a Gamma distribution for ISI values [53],  $p(H_1)$  is the probability of  $ISI < t_{rise}$ , calculated by

$$p(H_1) = \int_{x=0}^{0.2} \text{Gamma}(k, c) dx, \quad (26)$$

were the Gamma distribution is defined in Eq. 22. We determined the parameters  $k$  and  $c$  using the dataset from section 2.2.3 Extraction of Single and Double AP Evoked Fluorescence Transients. Since this dataset was obtained from anesthetized mice (which include very large ISIs not observed in awake state), we used only ISI values less than 1 ms for estimating biologically plausible values for the Gamma distribution in awake mice. Fitting Gamma distributions to the ISI values of individual neurons, we estimated a mean

value of  $k = 1$  and  $c = 0.2$ . Substituting these values in Eq. 26 and considering a high detection threshold of  $P_D = 0.99$  result in  $P_F = 0.017$ ; the same values of  $P_D$  and  $P_F$  are used for analyzing the resolution limits of GCaMP6s and GCaMP6f.

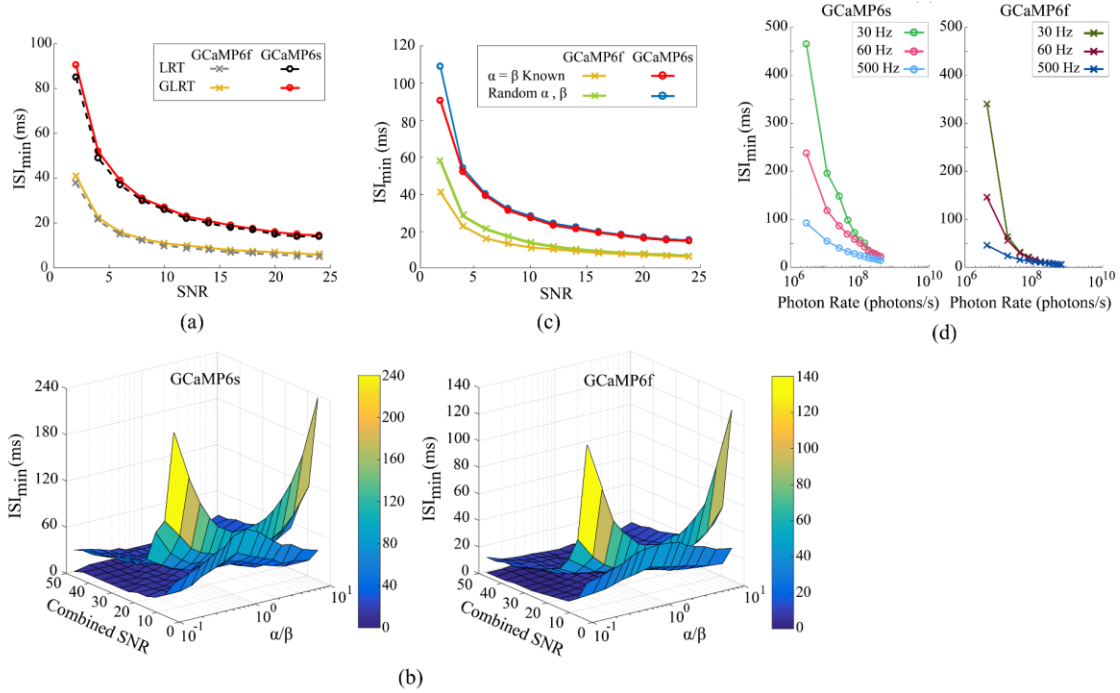
It is illuminating to see how well the GLRT detector performs compared to the best optimal detector (LRT), in which all the parameters are known. Figure 3(a) shows the smallest detectable ISI ( $ISI_{\min}$ ) of both sensors for the symmetrically located spikes with equal known amplitudes versus SNR for an AOD scanner operating at  $f_s = 500$  Hz. The results were obtained by generating receiver operating characteristic (ROC) curves from 2000 Monte-Carlo simulations at each SNR and ISI sampled with 0.5 ms spacing.  $ISI_{\min}$  was determined by the smallest ISI value for which the corresponding ROC curve satisfied  $P_D = 0.99$  and  $P_F = 0.017$ . Comparing the two detectors of each sensor, Figure 3(a) suggests that GLRT performs very close to the optimal detector. It also shows that we can accurately resolve ISI values much smaller than the fluorescence waveforms' rise times ( $t_{rise} \cong 200$  ms and 45 ms for GCaMP6s and GCaMP6f, respectively). At SNR = 3 (SNR = 2) for GCaMP6s (GCaMP6f), the detector distinguishes fluorescence waveforms with ISI as small as about 60 ms (40 ms).

Figure 3(b) compares two cases of the known amplitudes, namely,  $d_1 = d_2$  and  $\alpha d_1 = \beta d_2$ , for different combined SNR levels (i.e., sum of the two transients' SNRs) and amplitude ratios between the waveforms. The  $\alpha \neq \beta$  case gives better detection performance under the  $d_1 = d_2$  condition, suggesting that at a fixed SNR level, we can resolve smaller ISIs compared to the equal amplitudes case. This result was also reported

in [41] for a symmetric PSF. As explained in section 2.2.2.1 Spikes with known amplitudes, for the case of  $\alpha \neq \beta$  with  $d_1 = d_2$ , the  $H_0$  hypothesis is not located in the most challenging distance between two signals of the  $H_1$  hypothesis, making the detection problem easier. However, when the test is conducted according to Eq. 10 the  $\alpha \neq \beta$  case is a more challenging problem compared to  $\alpha = \beta$ . That is, with the same SNR level, the detector can resolve a larger ISI. This result emphasizes the importance of the  $H_0$  hypothesis in the performance of the detector.

For the case of unknown amplitudes with prior probability distribution, as explained in section 2.2.2.2 Spikes with random amplitudes, the performance of the detector is characterized by averaging  $P_D$  and  $P_F$ . Since a closed form expression is not available for  $P_D$  and  $P_F$  relating them to  $\alpha$  and  $\beta$ , Monte-Carlo simulation with  $f_s = 500$  Hz was used to numerically solve the problem. At each SNR value, 200 independent values of  $\alpha$  and  $\beta$  were drawn from their prior distribution. For each draw at each SNR and ISI sampled with 0.5 ms spacing, 2000 simulations were executed, and the results are shown in Figure 3(c). Comparing the result of this problem with the known amplitude case, we note that the prior knowledge about the amplitudes in the random case has resulted in a performance very close to the known case, with the latter slightly outperforming the former especially at the low SNR = 2. In all cases, the utilized detector can distinguish the presence of two spikes at ISIs much smaller than the fluorescence waveforms' rise times. At the SNR levels of the GCaMP6s and GCaMP6f datasets (SNR = 3 and 2, respectively),

the detector for the general case of random amplitudes, on average, detects two fluorescence waveforms that are about 70 ms and 60 ms apart.



**Figure 3: The smallest detectable ISI ( $ISI_{\min}$ ) determined the resolution limit.  $ISI_{\min}$  smaller than the fluorescence waveform’s rise time can be achieved under certain experimental conditions. (a) GLRT achieves similar  $ISI_{\min}$  values to LRT. (b)  $ISI_{\min}$  versus combined SNR and ratio between signal amplitudes for (lower curves)  $d_1 = d_2$  and (upper curves)  $\alpha d_1 = \beta d_2$ . (c) Prior knowledge about the probability distribution of randomly distributed amplitudes results in similar detection performance to the equal known amplitude case. (d)  $ISI_{\min}$  calculated for different recording speeds. All results (c)-(d) were obtained with 2000 Monte-Carlo simulation and at detection performance point of  $P_D = 0.99$  and  $P_F = 0.017$ . Figure adapted from [26].**

We compare the detection performance of different recording speeds under equal dwell time and baseline photon emission rates in Figure 3(d). Results from this analysis indicate that higher recording speeds can resolve significantly smaller ISI values for both calcium sensors at low photon emission rates. Nonetheless, experimentalists equipped

with a conventional recording system can attain resolution limits smaller than the fluorescence waveform's rise time by imaging a smaller field-of-view and thus, increasing dwell time and SNR. Overall, when designing experiments, sensor properties, SNR, and frame rate should all be considered to achieve the desirable spike detection performance.

### 2.3.2.2 Detector performance on experimental dataset

We applied the formulated detector under the unknown amplitudes case on the experimental data described in section 2.2.3 Extraction of Single and Double AP Evoked Fluorescence Transients. The data extraction pipeline resulted in 82 and 258 two spike samples for GCaMP6s and GCaMP6f expressing neurons, respectively. In analyzing experimental data, the test origin needs to be determined first. This can be done by finding the maximum point of cross-correlation between individual signals and  $s_0(t_k)$ . To avoid erroneous calculation of the time origin due to noise, we used the true spike times to center the extracted signals on  $t = 0$ .

We determined the detection threshold based on a desired value for  $P_F$  common between all SNR values. This can be done because the probability distribution of the log-likelihood ratio under the  $H_0$  hypothesis is independent from the true values of parameters defining the model under  $H_0$  [27]. As an illustrative example, we performed the detection problem by setting  $P_F = 0.3$ . Figure 4(a) illustrates examples of one and two spike fluorescent signals that were either correctly or incorrectly labeled by the detector. Fixing the detection threshold at the desired  $P_F$  level, different theoretical  $P_D$  values are derived from the ROC curves of 2000 simulated data for each different SNR and ISI pair

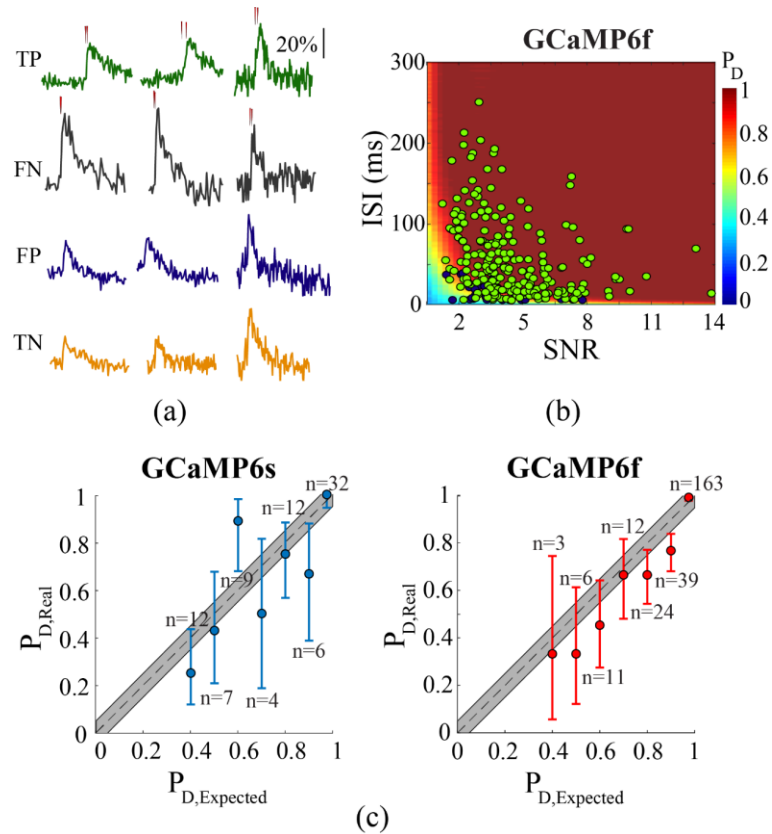
values (Figure 4(b) illustrates the case for GCaMP6f). The detector achieved total detection rates of 0.74 and 0.87 for GCaMP6s and GCaMP6f datasets, respectively. It also resulted 0.26 in (GCaMP6s) and 0.37 (GCaMP6f) total false positive rates, which are approximately the expected values from setting  $P_F = 0.3$ .

To further compare the detector's expected performance through theoretical analysis to the observed performance on experimental data, we took the following steps. First, we grouped the two spike data points based on their theoretical  $P_D$  values ( $P_{D,Expected}$ ). We discretized  $P_D$  values by rounding to obtain the sample groups. Next, real  $P_D$  value ( $P_{D,Real}$ ) for each group was calculated as the percentage of samples correctly detected as two spikes in each group. We utilized the binomial confidence interval to assess  $P_{D,Real}$  values [54]. Since the number of samples in each group was relatively small, we used the 68% confidence interval corresponding to data within one standard deviation of the mean to assess whether our detector attained performance close to the theoretically predicted performance. Our analysis revealed that the detector's detection performance on experimental data was indeed close to that predicted in theory (Figure 4(c)), as the  $P_{D,Expected}$  values fall in the confidence intervals of  $P_{D,Real}$ .

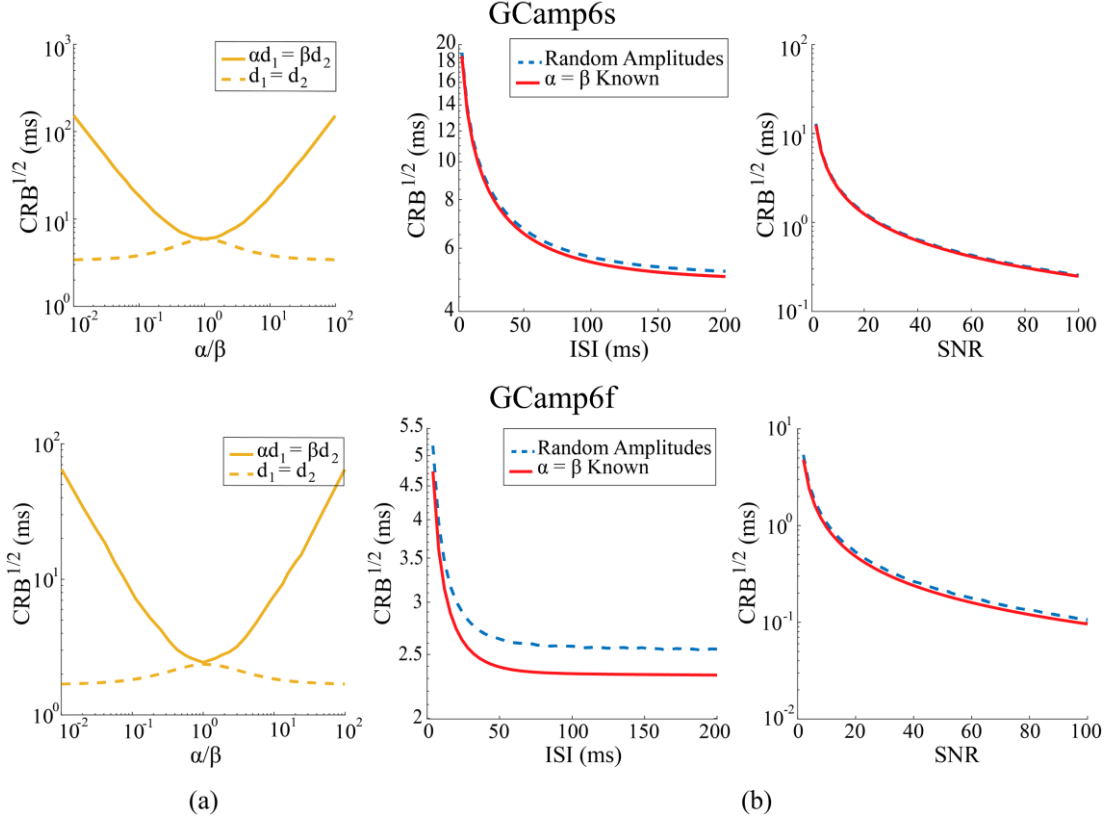
### 2.3.3 Prior Knowledge about Signal Amplitudes Yields Theoretically Equal ISI Estimation Performance to the Known Case

In this subsection, we present the CRB and HCRB for the known and the random amplitude cases, respectively. Figure 5(a) illustrates the effect of amplitude ratios on  $CRB^{1/2}$  for the  $\alpha \neq \beta$  case under the two conditions,  $d_1 = d_2$  and  $\alpha d_1 = \beta d_2$ , with fixed ISI = 60

ms and combined SNR = 8. As the ratio between the amplitudes diverges from one, CRB gets larger for the  $\alpha d_1 = \beta d_2$  case, whereas it decreases in  $d_1 = d_2$ . This result is similar to the result in Figure 3, where we emphasized on the effect of defining the  $H_0$  hypothesis.



**Figure 4: Two spike detection results from the experimental dataset.**(a) Examples of true positive (TP), false negative (FN), false positive (FP), and true negative (TN) from the GCaMP6f dataset. Vertical red lines correspond to spike times in the two spike cases. (b) GCaMP6f two spike samples overlaid on heat map of  $P_D$  versus SNR and ISI at a fixed  $P_F$ . Green and navy circles denote TP and FN samples, respectively ( $n = 258$  samples). The detector obtained 0.87 TP rate and 0.37 FP rate. (c) The detector approximately achieves the expected performance calculated through theoretical analysis. Error bars indicate 68% confidence intervals. Gray shaded areas denote discretized intervals. Number of samples in each interval is written along the corresponding interval. All results are obtained at fixed  $P_F = 0.3$ . Figure taken from [26]. © [2018] IEEE



**Figure 5: Lower bounds on ISI estimation at  $f_s = 500$  Hz for GCaMP6s and GCaMP6f. (a)  $CRB^{1/2}$  for two cases of known and unequal amplitudes versus the amplitude ratio at combined SNR = 8 and ISI = 60 ms. Estimating ISI from two unequally bright transients that are symmetrically located around the test origin gives better precision. (b)  $CRB^{1/2}$  and  $HCRB^{1/2}$  for the case of known and random amplitude cases, respectively, versus (left) ISI at combined SNR = 8, and (right) SNR at ISI = 60 ms ( $\alpha d_1 = \beta d_2$ ). An optimized ISI estimator with a random but known prior distribution about the amplitudes asymptotically performs similar to an optimized unbiased estimator of ISI with known  $\alpha = \beta$ . Figure adapted from [26].**

We complete the derivation of the HCRB described in section 2.2.4.2 Hybrid CRB for random amplitude signals by calculating  $\mathbf{I}_P$  according to Eq. 21. Based on the i.i.d assumption of  $\alpha$  and  $\beta$ , and their independence from ISI, only the second and third diagonal elements of  $\mathbf{I}_P$  corresponding to  $\alpha$  and  $\beta$  are non-zero and equal. Referring that  $\alpha, \beta \sim \text{Gamma}(k, c)$ , these two elements are derived as

$$\mathbf{I}_{P2,2} = \mathbf{I}_{P3,3} = -\mathbb{E}\left\{\frac{\partial^2 \ln p(\alpha)}{\partial \alpha^2}\right\} = \mathbb{E}\left\{\frac{k-1}{\alpha^2}\right\}. \quad (27)$$

Note that

$$x \triangleq 1/\alpha \sim \text{InvGamma}(k, c^{-1}). \quad (28)$$

with the probability distribution function defined as

$$f(x; k, c^{-1}) = \frac{c^{-k} x^{-k-1} \exp\left(-\frac{c^{-1}}{x}\right)}{\Gamma(k)}, \quad x > 0. \quad (29)$$

The mean and variance of this distribution are  $c^{-1}/(k-1)$  (for  $k>1$ ) and  $c^{-2}/[(k-1)^2(k-2)]$

(for  $k>2$ ), respectively [52]. Thus, we arrive at

$$\mathbf{I}_{P2,2} = \mathbf{I}_{P3,3} = -\mathbb{E}\left\{\frac{k-1}{\alpha^2}\right\} = (k-1) \left[ \text{var}\left(\frac{1}{\alpha}\right) + \left(\mathbb{E}\left\{\frac{1}{\alpha}\right\}\right)^2 \right] = \frac{c^{-2}}{k-2}. \quad (30)$$

for  $k>2$ . Thus,  $\mathbf{I}_P$  is attained as

$$\mathbf{I}_P = \begin{bmatrix} 0 & 0 & 0 \\ 0 & c^{-2}/(k-2) & 0 \\ 0 & 0 & c^{-2}/(k-2) \end{bmatrix}.$$

We compare the  $\text{HCRB}^{1/2}$  with  $\text{CRB}^{1/2}$  of the known and equal amplitude case at combined  $\text{SNR} = 8$  and  $\text{ISI} = 60$  ms in Figure 5(b); The two bounds are nearly identical (mean  $\pm$  std difference of  $0.2 \pm 0.05$  ms and  $0.24 \pm 0.04$  ms (right) and  $0.05 \pm 0.07$  ms and  $0.05 \pm 0.09$  ms (left) for GCamp6s and GCamp6f, respectively). We thus conclude that an optimized unbiased ISI estimator with known  $\alpha = \beta$  asymptotically performs similar to an optimized ISI estimator with a random but known prior distribution about the amplitudes.

### 2.3.4 Maximum Likelihood and Maximum a Posteriori Estimators Closely Approach the Theoretical Bounds

In this section, we compare the performance of ML and MAP estimators with their corresponding lower bounds. Figure 6(a) and (b) show the comparison of bias and standard deviation of ISI estimation for both GECIs (through 5000 Monte-Carlo simulations) to the  $\text{CRB}^{1/2}$  limit, assuming symmetrically located spikes with known amplitudes fixed at combined SNR = 20. Except for very small ISI values in Figure 6(a), the results show that the ML estimator is unbiased and its standard deviation is very close to the lower limit, emphasizing its ability to achieve the theoretically best possible precision. However, for small values of ISI in the  $\alpha = \beta$  problem, the standard deviation of ISI estimations becomes smaller than the lower limit. Note that in the  $\alpha = \beta$  problem, the maximization problem in Eq. 6 has two answers:  $\text{ISI} = d$  and  $\text{ISI} = -d$ . The ML estimator achieves asymptotic consistency and efficiency under certain conditions; one is that the maximum of Eq. 6 should be unique [55]. For large ISI values, where the two peaks are relatively far from each other, iterative optimization methods used to numerically solve the maximization problem converge to one of the two peaks depending on the starting point. Therefore, we may assume a “unique” peak at the local region around either one of the maximums where the consistency and efficiency properties hold. However, as ISI gets smaller and the precision of estimation decreases, observation noise will deviate the peak locations of the log-likelihood function towards  $\text{ISI} = 0$ . In the presence of such bias, the comparison of ML variance to CRB is theoretically invalid. We specify the boundary of the valid region for ML and CRB comparison based on the simulation results presented

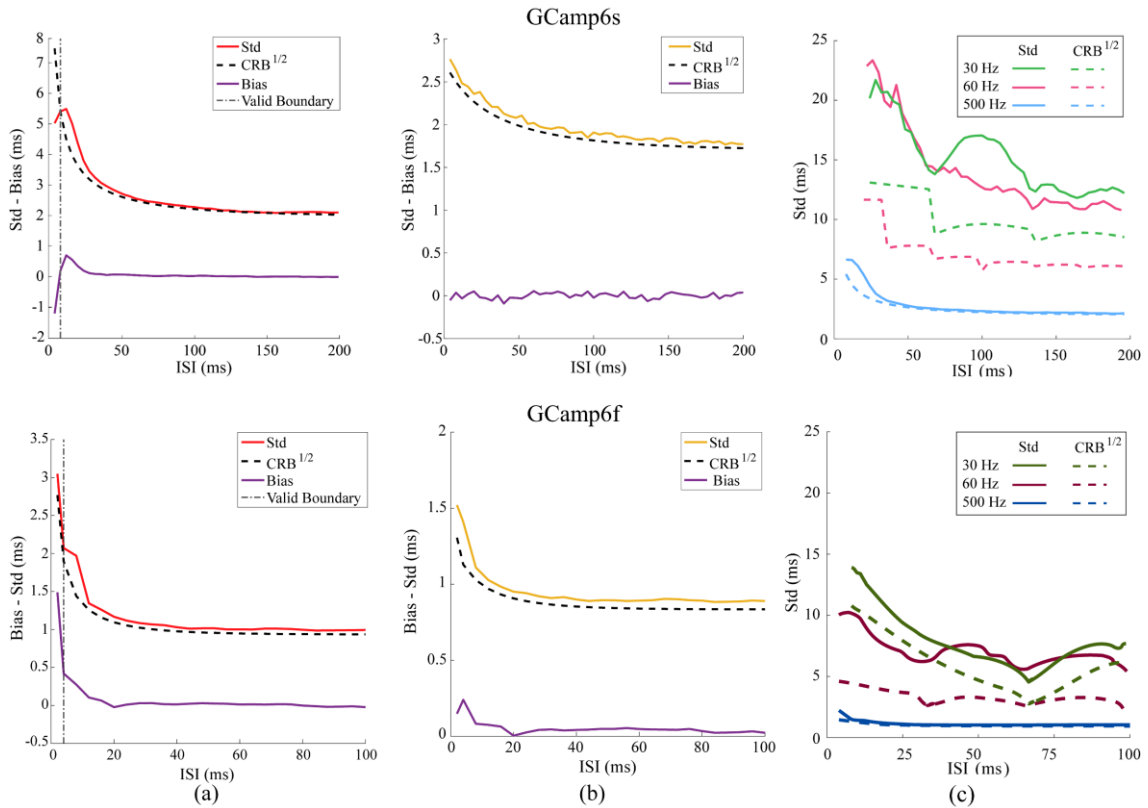
in Figure 7. The histograms of numerically calculated ISI values at combined SNR = 20 show that for small values of ISI, a discernible peak appears at ISI = 0. This results from the estimator being trapped around zero.

At ISI > 8 ms (GCaMP6s) and ISI > 4 ms (GCaMP6f) the peak at zero becomes less prominent peak height becomes smaller than half of the height at true value), reducing the bias. Therefore, we determine ISI = 8 ms and 4 ms as the boundary of the valid region of ML and CRB comparison for GCaMP6s and GCaMP6f with 500 Hz recording speed, respectively. The estimations to the left of these boundaries have considerable bias, therefore making the comparison of the standard deviation to the  $\text{CRB}^{1/2}$  invalid.

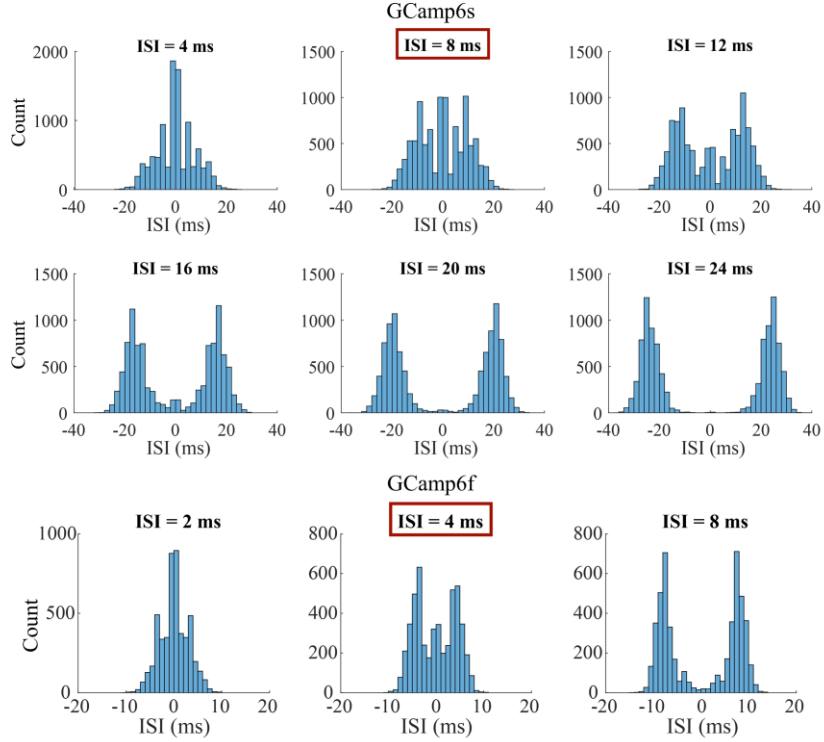
Figure 8 compares the root-mean-squared error (RMSE) of the GECI parameter estimations to the  $\text{HCRB}^{1/2}$  limits versus ISI for the case of random amplitudes with  $\alpha d_1 = \beta d_2$ , and combined SNR = 20 and  $f_s = 500$  Hz. The results suggest that the simultaneous ML and MAP estimation of ISI and the amplitudes achieves very close performance to the asymptotic limit, especially for  $\alpha$  and  $\beta$ . In the small ISI region, bias in the amplitude estimations towards  $\alpha = \beta$  leads to the previous problem of non-unique solution for ISI. Therefore, we included the valid boundary as in Figure 6(a) for completeness. The comparison of the results on the left of this line to the boundary is not valid.

Finally, we analyze the information theoretic lower bound for different recording systems. Under the equal amplitude case, Figure 6(c) compares the calculated standard deviation of ISI estimation through 5000 Monte-Carlo simulations to the lower bounds for both calcium sensors at the fixed combined SNR = 20. At the same SNR level, the CRB for

ISI estimation using higher recording rates is smaller compared to lower recording rates. More importantly, the very high 500 Hz recording speed comes very close to achieving its estimation lower bound, as can be seen from the small distance between the calculated standard deviation and the theoretical lower bound.



**Figure 6: ML estimators nearly achieved the information-theoretic bounds. Results obtained from 5000 Monte-Carlo simulations at combined SNR = 20. Standard deviation (std) and bias of ISI estimation compared to  $\text{CRB}^{1/2}$  for known values of (a)  $\alpha = \beta$  and (b)  $3\alpha = \beta$  with  $d_1 = d_2$  at  $f_s = 500$  Hz. (c) Std of estimating ISI from simulated dataset compared to  $\text{CRB}^{1/2}$  for the case of equal amplitudes with different recording rates. Only valid regions are depicted. Figure adapted from [26].**



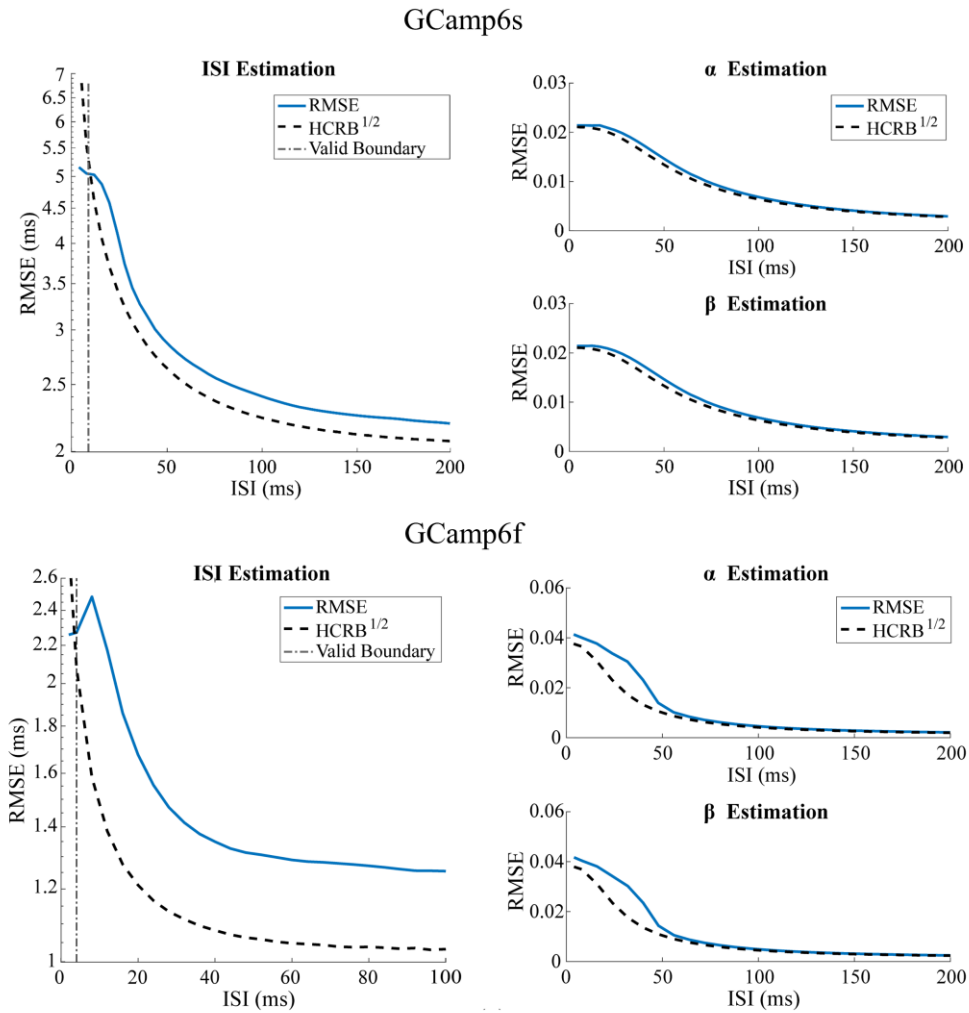
**Figure 7: Histograms of experimentally calculated ISI values in the  $\alpha = \beta$  known case and combined SNR = 20. Discernible peaks at ISI = 0 introduce large bias in the estimation, making the comparison of ML and CRB invalid. At ISI > 8 ms (GCamp6s) and ISI > 4 ms (GCaMP6f) the peak at zero becomes less prominent.**

Figure taken from [26]. © [2018] IEEE

## 2.4 Discussion

Calcium sensor kinetics and SNR significantly impact spike detectability and precision of spike time estimation [12]. Thus, there is a need for understanding the theoretical resolution limits of detecting closely-timed neuronal spikes from fluorescence signals. Similar to studies that have shown resolutions beyond the Rayleigh Diffraction limit is possible in optical imaging systems [40-43, 56], our work showed that by using the statistical approach, attaining resolution finer than the peak time of the indicator is possible. While this result was expected from a previous algorithm conducted on OGB-1

labeled neurons which assumed uniform calcium spike responses [5], our detector achieved equal performance considering randomly distributed calcium responses. The latter scenario better matches the true, stochastic response of calcium indicators during live animal experiments. The CRB lower bounds on the variance of ISI estimation further verified the results of the detection framework.



**Figure 8: MAP estimators nearly achieve the information-theoretic bounds. (left) ISI and (right)  $\alpha$  and  $\beta$  estimations compared to their HCRB $^{1/2}$ , with  $\alpha d_1 = \beta d_2$  at  $f_s = 500$  Hz. Dashed-dot vertical line shows the smallest ISI for which the CRB comparison is valid. Figure adapted from [26].**

Our detection theoretic framework assumed no definite knowledge about the peak value of a single spike, which is beneficial for modeling experiments with no ground truth available. This was a particularly challenging case since the peak amplitude of the single spike in the  $H_0$  hypothesis is comparable to the peak amplitude of the signal generated by two spikes in the  $H_1$  hypothesis. Thus, a simple decision between  $H_0$  and  $H_1$  based on amplitude alone, especially in low SNRs, cannot provide accurate results. We showed that utilizing the signal's temporal information, as modeled through  $s_0(t; \theta_0)$  and  $s_1(t; \theta_1)$ , enabled accurate detection.

The resolution limits and estimation bounds were estimated based on a set of experimentally derived parameters. We determined the detection criteria, i.e.  $P_D$  and  $P_F$ , by relating them through the prior probabilities of the two hypotheses, which were derived using the available dataset with ground truth spike times. The prior probabilities derived in our work are applicable to this specific data and need to be recomputed for any new experiment. In general, accurate information about the spiking behavior of the neurons might not be available. In such events, experimentalists can use any desirable values for  $P_D$  and  $P_F$  to derive the resolution limits of detecting temporally overlapping fluorescence waveforms. A good performing detector is one with very high  $P_D$  (usually above 0.9) and low  $P_F$  (such as 0.01). In general, a very high  $P_D$  along with a very low  $P_F$  value will make the detection of two spikes a harder problem, resulting in larger resolution limits (i.e.  $ISI_{\min}$ ).

Our model was based on Poisson statistics of the signals, which is generally true for shot-noise limited recordings. Importantly, we have demonstrated that our formulated detector performs as expected on experimentally obtained datasets. However, under certain conditions this assumption can be violated. For example, some signal extraction methods are based on the weighted average of multiple pixel values, generating signals that are not purely Poissonian. Another case is when neuropil contamination is removed by subtracting the average pixel values around the neuron soma. Nevertheless, our formalism should allow incorporation of other noise models in future work.

We have assumed linear relationship between calcium dynamics and fluorescence response. In general, this relationship is non-linear and sensor saturation occurs at very high firing rates. This effect is especially pronounced for past generations of protein calcium sensors with high dissociation constant. For the case of GCaMP6 sensors considered here, which are currently the best GECIs due to their favorable properties, the fluorescence response is linear in the low spike regime [45]. Therefore, the non-linear dynamics and saturation assumptions are not necessary for the work presented here, which deals with one and two spike cases.

This work is the first step in the continuum research to utilize detection theoretic tools to set the optimal resolution limits for temporally overlapping fluorescence signals. Future work will extend the current framework to the more general case of more than two spikes. Such analysis should take into account the non-linearity and saturation effect [57, 58].

### 3. Active Neuron Segmentation from Two-Photon Calcium Imaging Recordings Using Deep Learning

Automated, fast, and reliable active neuron segmentation from calcium imaging recordings is a critical step in the analysis workflow for discovery of neuronal coding properties in real-time behavioral studies. The purpose of this chapter was to exploit the full spatio-temporal information in two-photon calcium imaging movies for segmenting active neurons. The method developed in this chapter was published in the *Proceedings of National Academy of Sciences* under the title “Fast and robust active neuron segmentation in two-photon calcium imaging using spatiotemporal deep learning” [59] and presented at the *40th International Conference of The IEEE Engineering in Medicine and Biology Society* under the title “Deep-learning based active neuron segmentation in two-photon calcium imaging”. The contents in this chapter, including texts, figures, and tables were mainly reproduced from these publications.

We proposed a three-dimensional convolutional neural network to identify and segment active neurons from two-photon calcium imaging data. By utilizing a variety of two-photon microscopy datasets, we showed that our method outperformed state-of-the-art techniques and was on a par with manual segmentation. Furthermore, we demonstrated that the network trained on data recorded at a specific cortical layer can be used to accurately segment active neurons from another layer with different neuron density. Finally, our work documented significant tabulation flaws in one of the most cited and active online scientific challenges in neuron segmentation.

### ***3.1 Introduction***

Fast, automatic processing of the large calcium imaging datasets is a critical yet challenging step for discovery of neuronal coding properties in behavioral studies. Often the investigators are interested in identifying a subset of active neurons from the large imaged population, further complicating the neuronal segmentation task. The subset of modulating, and thus active, neurons in many behavioral experiments carry the meaningful information for understanding the brain's coding characteristics. Automatic identification of active neurons from the imaging movies in high speed enables scientists to directly provide dynamic complex behavioral or neural stimulus to the subjects in real-time.

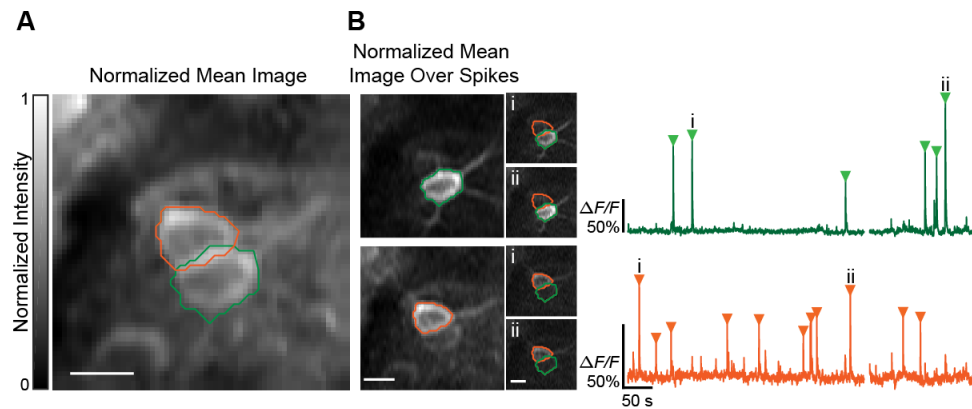
Recent efforts from several groups have produced automatic methods to detect and quantify neuronal activity in calcium imaging data. These methods span from unsupervised classic machine learning techniques [14, 16, 17, 19, 20, 60-65] to deep-learning based supervised algorithms [66, 67]. Among the former class of neuron segmentation algorithms are the popular methods of principal component and independent component analysis (PCA/ICA) [14], constrained non-negative matrix factorization (CNMF) [16], extension of CNMF to one-photon microscopy [65], and the more recent and faster version of CNMF, called OnACID [19], which is based on online dictionary learning. Recently, Giovannucci et al. [18] have improved the scalability of CNMF and extended OnACID with new initialization methods and a convolutional neural network (CNN), referred to as CaImAn Batch and CaImAn Online, respectively.

In general, the accuracy of assumptions in these model-based methods in characterizing the embedded patterns is a critical factor in the performance of such methods [68]. For example, CNMF models the background as a low-rank matrix, which might not capture the complex dynamic of the background in one-photon imaging recordings. To compensate for this background, Zhou et al. [65] incorporated an autoregressive model for the background components to process one-photon imaging data.

Deep learning can serve as an alternative to the above classic machine learning techniques. CNNs learn hierarchies of informative features for a specific task from labeled datasets [68]. Modern fully convolutional neural networks have become a staple for semantic image segmentation, providing an end-to-end solution for the pixel-to-pixel classification problem [69]. These networks are often more efficient compared to the traditional CNN-based segmentation approaches that label each pixel of an image based on the local intensity values [69].

A few recent approaches have utilized CNNs to segment neurons from two-dimensional (2D) images for subsequent temporal analysis. These methods treat multiple frames of imaging data as either additional channels [66] or one image averaged from all frames (the “mean image”) [67]. One example of this class of CNN-based methods is the method of Apthorpe et al. [66], which applies 2D kernels to individual frames and aggregates temporal information with a temporal max-pooling layer in the higher levels of the network. While the performance was not significantly different from a similar network that only processed the mean image, this CNN method outperformed PCA/ICA.

More recently, based on the fully convolutional UNet [70], Klibisz et al. [67] developed the UNet2DS method that segments neurons from the mean image. In general, these methods are suboptimal for differentiating active from non-active neurons due to the loss of temporal dynamics when summarizing temporally collected images into a mean image. Similarly, sparsely firing neurons may appear at unidentifiable contrasts compared to the background after undergoing averaging to the mean image. Lastly, 2D segmentation of mean images has difficulty in delineating the neuron boundaries between overlapping neurons that independently fire in time (Figure 9).



**Figure 9: Overlapping neurons complicate active neuron segmentation. (A) Neurons can have overlapping regions due to the projection of a 3D volume onto a 2D imaging plane. (B) The temporal evolution of neuron intensities provides important information for accurate segmentation of overlapping cases, which is exploited by our proposed method. The time-series in green and orange correspond to neurons outlined with matching colors. Images in the middle panel show the recorded data at the marked time-points, and the images in the left panel are the normalized mean images of frames corresponding to each neuron’s active time-interval (defined as 0.5 seconds after the marked spike times). Scale bars: 10  $\mu\text{m}$ . Figure taken from [59].**

Three-dimensional (3D) CNN architectures could be superior to 2D segmentation networks as they have the advantage of incorporating temporal information into an end-to-end learning process [71]. Compared to methods that process 2D images,

spatiotemporal methods can provide more accurate results in identifying sparsely spiking and overlapping neurons, but are also computationally more challenging [16]. Compared to iterative methods such as CNMF, a 3D CNN architecture could produce high computational efficiency for long-duration, large-scale recordings. 3D CNNs have already been impactful in other video [71, 72] and volumetric biomedical [73-75] data analyses.

A critical factor prohibiting development and accurate assessment of such novel learning-based techniques (e.g. 3D CNNs) is the absence of a comprehensive public dataset with accurate gold-standard ground truth markings. Indeed, the Allen Brain Observatory (ABO) (<http://observatory.brain-map.org/visualcoding>) and the Neurofinder challenge (<https://github.com/codeneuro/neurofinder>) have provided invaluable online resources in the form of diverse datasets spanning multiple brain areas. We demonstrate that existing markings that accompany these datasets contain significant errors, further complicating algorithm development and assessment. Like many other medical imaging modalities that lack empirically driven ground truth, human expert markings could serve as the gold-standard. In such situations, the agreement between multiple expert human graders has traditionally determined the practical upper bound for accuracy. No automated algorithm to-date is shown to be closer in accuracy to the markings of an expert human grader than another experienced grader.

Here we present a novel CNN-based method with spatiotemporal convolutional layers to segment active neurons from two-photon calcium imaging data. To train and validate the performance of this algorithm, we utilize online datasets from the ABO and

Neurofinder challenge. Since we show that the original manual markings that accompany these datasets are imperfect, we carefully manually-relabel active neurons in these datasets. We compare the performance of our network with other state-of-the-art neuron segmentation methods on these datasets. The results indicate that our trained network is fast, superior to other methods, and achieves human accuracy. To demonstrate the generalizability of our method, we show that the network trained on data recorded at a specific cortical layer from the ABO dataset can also accurately segment active neurons from other layers and cortical regions of the mouse brain with different neuron types and densities. We demonstrate that adding region-specific recordings to the ABO training set significantly improves the performance of our method. To promote future advancement of neuron segmentation algorithms, we have provided the manual markings, source code for all developed algorithms, and weights of the trained networks as an open-source software package (<https://github.com/soltanianzadeh/STNeuroNet>).

## ***3.2 Methods***

### **3.2.1 Allen Brain Observatory Dataset and Labeling**

This dataset consists of two-photon recordings from neurons across different layers and areas of the mouse visual cortex. Transgenic Cre-line mice drove expression of the genetically encoded calcium indicator GCaMP6f. We selected a subset of the entire recordings in our work. This dataset included the first 12 minutes and 51 seconds of recordings from 275  $\mu\text{m}$  and 175  $\mu\text{m}$  deep in the primary visual cortex (VISp) of ten mice per cortical depth. Table 4 shows the correspondence between the mouse lines and videos

used in our work. We used the recordings at 275  $\mu\text{m}$  deep in the cortex to compare different algorithms and the recordings at 175  $\mu\text{m}$  to assess the generalizability of all methods.

**Table 4: Description of data used from the Allen Brain Observatory. All data are from the primary visual cortex.**

Cortical Layer	Transgenic Line	Experiment ID
275 $\mu\text{m}$	Cux2-CreERT2-Cre	539670003, 531006860 501574836, 501484643 503109347, 534691284 502608215, 501729039
	Rorb-IRES2-Cre	510214538, 527048992
175 $\mu\text{m}$	Cux2-CreERT2-Cre	501704220, 501836392 510514474, 504637623 501271265, 502115959 502205092, 510517131
	Emx1-IRES-Cre	540684467, 545446482

The data was previously corrected for motion and had an accompanying set of automatically identified neurons that were not manually inspected [76]. We used these automatically detected neurons as initializations for our manual labeling. We developed a custom software with graphical user interface (GUI) in MATLAB 2017b (Mathworks, Natick, MA) that allowed our graders to add to the initial set by drawing along the boundary of newly found neurons (phase 1) and to dismiss wrongly segmented neurons that do not correspond to the soma of an active neuron (phase 2). In phase 1, the GUI provided simultaneous visualization of the video overlaid with segmented neurons' masks on two separate panels. On one panel, background corrected video and in the other

panel a summary image of choice (mean, max-projected, or correlation image, defined as the mean correlation value between each pixel with its 4-connected neighborhood pixels) were displayed. In phase 2, the GUI showed the zoomed-in region of the video for each segmented neuron in three panels, which included the background corrected video, the mean image, and the  $\Delta F/F$  trace of the average pixel intensities within the neuron's mask. Graders used the following criteria to label each marked mask as neuron: (1) the marked area had a bright ring with a dark center that changed brightness during the recording, or (2) the area was circular and had a size expected from a neuron (10-20  $\mu\text{m}$  in diameter) that changed brightness during the recording. Criterion 1 filters for nuclear-exported protein calcium sensors used by the ABO investigators, while criterion 2 filters for spatio-temporal features of neuron somas that have calcium activity transients.

Two graders independently corrected the markings of the ABO dataset. Matching marks from the two graders were labeled as true neurons, whereas disagreements were reevaluated by the senior grader (grader #1). This grader, blind to the identity of the non-matching masks (meaning who marked it), used the phase 2 of our GUI to assess all disagreements and label the masks as neuron or not a neuron. The set of masks marked by both graders and the set of masks that corresponded to active neurons from the disagreement set comprised the final ground truth (GT) masks. We created spatio-temporal neuron labels for training by extracting the neurons' active time-intervals. We first separated traces of overlapping neurons using the linear regression approach of [76]. Using the extracted time-series for each neuron, we removed neuropil signal, scaled by

factor of 0.7 [45], and removed any remaining background fluctuation using a 60 s moving-median filter. For each neuron, we defined the neuropil signal as the average fluorescence value in an annulus of 5  $\mu\text{m}$  around the neuron mask, from which we excluded pixels that belonged to other neurons. We found activity-evoked calcium transients with high detection fidelity (see section 3.2.8 Spike Detection and Discriminability Index). We considered neurons as active until 0.5 seconds after the detected spike times, equal to 3.5 times the half-decay time of spike-evoked GCaMP6f fluorescence signals reported in [45].

### 3.2.2 Neurofinder Dataset and Labeling

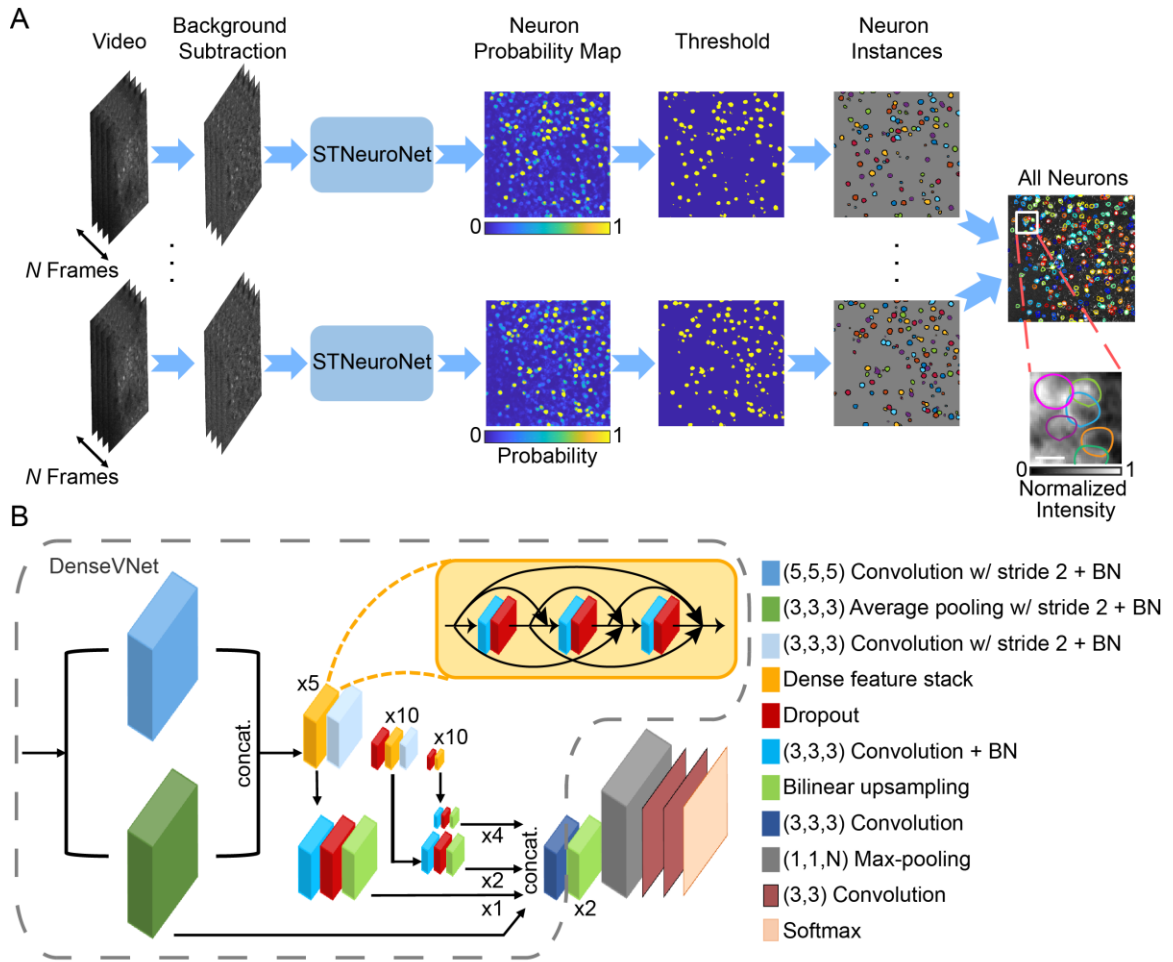
The Neurofinder challenge consists of nineteen training and nine testing two-photon calcium imaging datasets acquired and annotated by four different labs. These datasets are diverse: they reported activity from different cortical and subcortical brain regions and varied in imaging conditions such as excitation power and frame rate. The GT labels were available for the training sets, while they were held out for the test set.

The first dataset (called the 00 set) segmented neurons using fluorescently labeled anatomical markers, while others were either manually marked or curated with a semi-automatic method. Upon inspection of the fourth dataset (called the 03 set), we found that this dataset was labelled based on anatomical factors. We excluded the first and fourth sets from the comparison in the Results section because the activity-independent marking of these datasets is incompatible for assessing active neuron segmentation methods. The remaining datasets referred to as 01, 02, and 04 each had two training videos. Similar to

ABO, we created spatio-temporal labels for the Neurofinder training set by detecting neuronal spikes that satisfied the minimum required  $d'$  (see section 3.2.8 Spike Detection and Discriminability Index), which we iteratively reduced down to  $d' = 0.5$  if a spike was not identified.

### 3.2.3 Proposed Active Neuron Segmentation Method

The key feature of our active neuron segmentation framework (Figure 10A) was a new 3D CNN architecture, which we named SpatioTemporal NeuroNet (STNeuroNet) (Figure 10B). The 3D convolutional layers in STNeuroNet extracted local spatiotemporal information that capture the temporal dynamics of the input recording. STNeuroNet consisted of downsampling, upsampling, convolutional skip connections, and temporal max-pooling components that predict neuron masks based on spatiotemporal context of the input recording. After initial background compensation of individual movie frames, STNeuroNet processed sequences of short temporal batches of  $N = 120$  frames and output a 2D probability map of active neurons for each batch. We then applied an optimal threshold to the neuron probability maps and automatically separated high probability regions into individual neuron instances. Lastly, the final set of unique active neurons for the entire recording was determined by eliminating duplicate masks of the same neurons that were identified in different temporal intervals of the video. These steps are discussed in detail in the following sections.



**Figure 10: Schematic for the proposed spatiotemporal deep learning-based segmentation of active neurons in two-photon calcium videos.**(A) After removing background non-uniformity of each video batch ( $N = 120$  frames), STNeuroNet predicts the neuron probability map. We identify neuron instances in the binarized probability maps from multiple temporal batches, which we then fuse into the final set of active neurons for the entire video. The right inset is the mean image of the region enclosed by the white box. Scale bar:  $10\ \mu\text{m}$ . (B) STNeuroNet architecture details. Numbers on top of the dense feature stacks indicate the number of convolutional layers involved, and numbers for the bilinear upsampling blocks indicate the upsampling factor. BN: Batch normalization. Figure taken from [59].

### 3.2.4 Image Processing Steps

All data used in our work were previously registered. We first cropped the boundary region of the data to remove black borders introduced in the registration processes (10  $\mu\text{m}$  in each direction for the ABO data and 4-50  $\mu\text{m}$  for the Neurofinder data). To increase SNR, reduce the computational complexity, and allow utilization of the trained network for future data with different recording speeds, we temporally binned ABO and Neurofinder videos to 6 Hz and 3 Hz videos (lowest frame rate among the five datasets in the Neurofinder challenge), respectively. We performed temporal binning by combining a set of consecutive frames into one frame via summation. We then corrected for non-uniform background illumination using homomorphic filtering [77] on each frame of the video. We formulated a high-pass filter by subtracting a low-pass Gaussian filter with standard deviation of  $0.04 \mu\text{m}^{-1}$  from 1. Then, we normalized the intensity of each video by dividing by its overall standard deviation.

### 3.2.5 Neural Network Architecture

Much like action recognition from videos, active neuron segmentation requires capturing context from multiple frames. This motivated us to utilize 3D convolutional layers in our deep learning network. 3D convolutional layers extract local spatiotemporal information that capture the temporal dynamics of the input recording. We used the DenseVNet [78] as the backbone for our STNeuroNet network. DenseVNet is composed of downsampling, upsampling, and skip connection components (Figure 10B). Each encoder stage of is a dense feature stack. In the encoder path, strided convolutional layers

reduce the dimensionality of the input feature map and connect dense feature stacks. Single convolutional layers in the skip connections, followed by bilinear upsampling, transform the output feature maps from each stage of the encoder path to the original image size [78]. All convolutional layers in DenseVNet perform 3D convolutions, use the rectified linear unit (ReLU) non-linearity as the activation function, and consist of batch normalization [79] and dropout [80] with probability of 0.5 (except the last layer). Unlike the original implementation of the network, we did not use spatial priors, dilated convolutions, and batch-wise spatial dropout, as these did not have a significant effect on the final results reported in the original paper [78].

We made the following two modifications to DenseVNet for our application: (1) we changed the last convolutional layer of DenseVNet to have ten output channels instead of the number of classes, and (2) we added a temporal max-pooling layer to the upsampled features, followed by a 2D convolutional layer with ten  $3\times 3$  kernels, and a final convolutional layer with two  $3\times 3$  kernels to the output of DenseVNet. The temporal max-pooling layer summarizes the extracted temporal feature maps, greatly increasing the speed of the training process and reducing inference time by reducing the number of output predictions (2D predictions instead of 3D predictions). This step is important for high-speed network validation and low-latency inference. The last convolutional layer computes two feature maps for the background and neuron classes. We applied Softmax to each pixel of the final feature maps to transform them into probability maps. We used the Dice-loss objective function [75] during training, defined as

$$\text{Dice - loss} = 1 - \frac{2 \sum_{i=1}^N p_i q_i}{\sum_{i=1}^N p_i^2 + \sum_{i=1}^N q_i^2}, \quad (1)$$

where the summation is over  $N$ , the total number of pixels, and  $p_i$  and  $q_i$  are the Softmax output and GT label for pixel  $i$ , respectively. The Dice-loss is suitable for binary segmentation problems and handles highly unbalanced classes without the need for sample re-weighting [75].

### 3.2.6 Training the Network

To create a large set of training samples, we cropped smaller windows of size  $144 \times 144 \times 120$  voxels from the rotated ( $0^\circ$ ,  $90^\circ$  and  $180^\circ$ ) training videos and GT markings and applied random flips during training. We performed cropping using a spatiotemporal sliding window process with 75% overlap between adjacent windows. Within this large set of samples, we kept samples that contained at least one active neuron in the selected 120 frames time-interval. We trained the networks using sample-level whitening, defined as

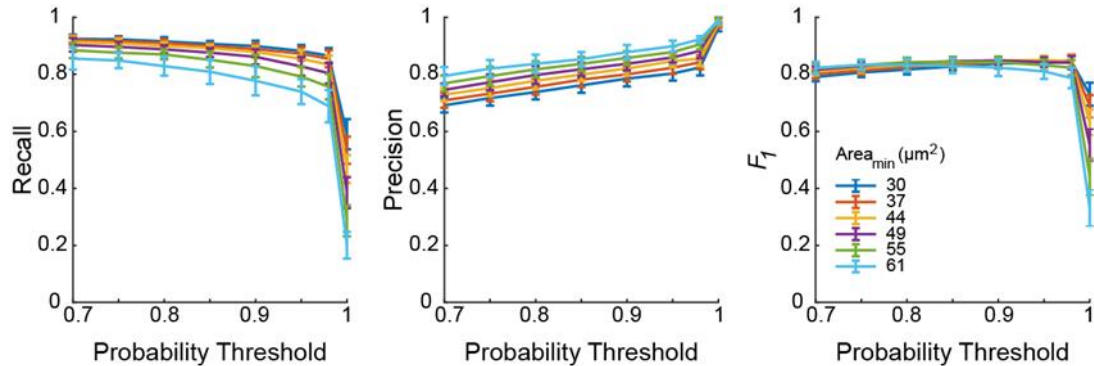
$$\frac{I - \text{mean}(I)}{\text{std}(I)}, \quad (2)$$

where  $I$  is the 3D input sample to the network. We used the Adam optimizer [81] with learning rate of 0.0005 and mini-batch size 3. We trained the ABO and Neurofinder networks for at least 35,000 iterations, or until the loss function converged (maximum 40,000 iterations). To optimally utilize our labeled dataset yet strictly separate training and testing datasets, we used leave-one-out cross-validation to assess the performance for detection and segmentation of active neurons.

### 3.2.7 Post-processing

#### 3.2.7.1 Binarizing neuron probability maps

We used the entire spatial extent of video frames at test time to estimate the neuron probability maps, which we processed to isolate individual neurons. We processed video frames in non-overlapping batches of  $N = 120$  frames, equal to the number of frames used during training. We binarized the probability maps by applying the optimal threshold that yielded the highest mean  $F_1$  score on the training set (Figure 11). We then separated potential overlapping active neurons from each binarized map and removed small regions. Finally, we aggregated all identified active neuron masks from different time-intervals to obtain the segmentation masks for the entire recording. These steps are described in detail in the following subsections.



**Figure 11: The optimal thresholds in the post-processing step of our algorithm are determined through leave-one-out cross-validation. Example results of recall, precision, and  $F_1$  scores by applying different levels of probability and minimum area thresholds to the training videos. Figure taken from [59].**

### 3.2.7.2 Instance segmentation

The temporal max-pooling layer in our network merges overlapping active neurons in the segmentation mask. To separate these neurons, we used the watershed algorithm [82]. We first calculated the distance transform image as the distance of each pixel to the nearest background pixel. We then applied the MATLAB *watershed* function to the distance transform of connected components which had an area greater than a predefined threshold, empirically set to the average neuron area ( $107.5 \mu\text{m}^2$  for ABO and  $100\text{-}200 \mu\text{m}^2$  for Neurofinder). After separating neuron instances, we discarded small segmented regions as background, with the minimum area determined to maximize the mean  $F_1$  score across the training set (Figure 11). Since the watershed algorithm alone cannot accurately determine neuron boundaries for overlapping cases, we used segmentation results from multiple temporal batches to yield the final neuron masks. This step is detailed in the following section.

### 3.2.7.3 Neuron fusion

Since STNeuroNet outputs a single 2D probability map of active neurons for the input time-interval, we processed two-photon video recordings in subsequent short temporal intervals to better resolve overlapping neurons. Unlike the approach of [66] which used the network predictions to find neuron locations, we used STNeuroNet's predictions to determine the final neuron masks. In each of these time-intervals, we identified and segmented active neurons. Because neurons may activate independently and spike in different times, we aggregated the segmentation results from all time-

intervals to attain the segmentation for the entire recording. Aggregation of neuron masks from multiple inferences corresponding to different time-intervals was done in two steps. First, we matched neurons between these segmentations to identify if the same neuron was segmented multiple times and kept the mask with the mean size. We used the distance between the masks' center of mass for this step. Masks with distance smaller than 4  $\mu\text{m}$  were identified as the same neuron. Second, we removed any mask that encompassed one or more neurons from other time-intervals. We removed any mask  $m_i$  that overlapped with mask  $m_j$  such that

$$\text{Normalized Overlap}(m_i, m_j) = \frac{|m_i \cap m_j|}{|m_j|} > \theta_p, \quad (3)$$

where  $\theta_p$  is the overlap threshold, which we empirically set to 0.75.

### 3.2.8 Spike Detection and Discriminability Index

Using tools from statistical detection theory [37, 83], we detected prominent spike-evoked fluorescence signals and quantified their detection fidelity. Specifically, we performed a matched filter approach with an exponentially decaying signal as the template ( $S$ ), with mean decay time of  $\tau$ , on the  $\Delta F/F$  traces to reduce the effect of noise on spike detection [37]:

$$L = F_0 \sum_{i=1}^n [-S_i + (1 + (\Delta F/F)_i) \ln(1 + S_i)], \quad (4)$$

in which the summation is over a sliding window of length  $n$ , and  $F_0$  is the baseline photon-rate. Using the relationship between the mean decay time  $\tau$  and half-decay time

$\tau_{1/2}$  as

$$\tau = \frac{\tau_{1/2}}{\ln(2)}, \quad (5)$$

we used 0.8 s and 0.2 s as the value of  $\tau$  for GCaMP6s and GCaMP6f data in  $S$ , respectively.

We detected spikes as local-maxima time points in a 1 s window of the filtered signal ( $L$ ) that passed a predefined threshold of  $\gamma$ :

$$\gamma = \mu + \sigma\Phi^{-1}(P_N), \quad (6)$$

which was determined by the tolerable probability of false-negative ( $P_N$ ) and the mean ( $\mu$ ) and standard deviation ( $\sigma$ ) of the distribution of  $L$  under the hypothesis of a spike having occurred [37]. In the above equation,  $\Phi^{-1}(\cdot)$  is the inverse of the standard Gaussian cumulative distribution function [37].

We further narrowed down the true spikes using the discriminability index,  $d'$ , which characterizes the detection fidelity by considering the amplitude and temporal dynamics of the fluorescence signals [37]. Higher values of  $d'$  provide higher spike detection probabilities and lower errors, with  $d' \geq 3$  achieving area under the receiver operating characteristic curve (a metric for spike detectability) greater than 0.98 [37]. We determined the minimum required detectability index ( $d'_{min}$ ) for labeling spikes with the aim of balancing the number of false-positive ( $P_F$ ) and false-negative ( $P_N$ ) errors [83]:

$$(f_s - \lambda)P_F = \lambda P_N, \quad (7)$$

$$d'_{min} = \Phi^{-1}(1 - P_N) - \Phi^{-1}(P_F) = \Phi^{-1}(1 - P_N) - \Phi^{-1}(P_N\lambda(f_s - \lambda)^{-1}). \quad (8)$$

In Eq. (7),  $f_s$  and  $\lambda$  denote the recording and neuron spike rates, respectively. For the ABO dataset, since the majority of mice were stationary during the visual stimulus behavior, we selected  $\lambda = 2.9$  spikes/s in accordance to previous experimentally obtained spike rates

during similar behaviors [84]. We then set a low  $P_N = 0.035$ , which corresponded to a spike detection threshold of  $d' = 3.6$  based on Eqs. (7)-(8). For the Neurofinder challenge, we used a lower threshold of  $d' = 1.7$  to compensate for the overall lower SNR of the data compared to the ABO dataset.

### 3.2.9 Quantification of Peak Signal-to-noise Ratio (PSNR)

To calculate the PSNR of neurons, we first separated traces of overlapping neurons using the linear regression approach of [76]. We then removed neuropil signal, scaled by factor of 0.7 [45], and removed any remaining background fluctuation using a 60 s moving-median filter. We then calculated the PSNR for neural traces as

$$\text{PSNR} = \frac{\Delta F_{\text{peak}}}{\sigma_n}, \quad (9)$$

where  $\Delta F_{\text{peak}}$  is the difference between the biggest spike value and the baseline value, and  $\sigma_n$  is the noise standard deviation calculated from non-active intervals of traces.

### 3.2.10 Evaluation Metrics

We evaluated segmentation methods by comparing their results with the manual GT labels. We assessed each algorithm by quantifying three neuron detection metrics: recall, precision, and F<sub>1</sub> score, defined as

$$\text{Recall} = \frac{N_{\text{TP}}}{N_{\text{GT}}}, \quad (10)$$

$$\text{Precision} = \frac{N_{\text{TP}}}{N_{\text{detected}}}, \quad (11)$$

$$F_1 = 2 \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}. \quad (12)$$

These quantities derive from the number of manually labelled neurons (ground truth neurons,  $N_{GT}$ ), number of detected neurons by the method ( $N_{detected}$ ), and number of true positive neurons ( $N_{TP}$ ). We used the Intersection-over-Union (IoU) metric along with the Hungarian algorithm to match masks between the GT labels and the detected masks [18].

The IoU for two binary masks,  $m_1$  and  $m_2$ , is defined as

$$\text{IoU}(m_1, m_2) = \frac{|m_1 \cap m_2|}{|m_1 \cup m_2|}. \quad (13)$$

We calculated the distance between any pair of masks from the GT ( $m_i^{GT}$ ) and the detected set ( $M_j$ ) as described by [18]:

$$\text{Dist}(m_i^{GT}, M_j) = \begin{cases} 1 - \text{IoU}(m_i^{GT}, M_j), & \text{IoU}(m_i^{GT}, M_j) \geq 0.5 \\ 0, & m_i^{GT} \subseteq M_j \text{ or } M_j \subseteq m_i^{GT} \\ \infty, & \text{otherwise.} \end{cases} \quad (14)$$

In the above equation, a distance of infinity denotes masks that are not matching due to their small IoU score. Next, we applied the Hungarian algorithm to solve the matching problem using the distance matrix defined in Eq. (14), yielding the set of true positive masks.

### 3.2.11 Speed Analysis

For each algorithm, we calculated the speed by dividing the number of frames by the processing time (excluding read and write times). For CaImAn Batch, we used all of the logical Cores of our CPU (28 threads) for parallel processing. For STNeuroNet and CaImAn online, we calculated an initialization-independent speed by disregarding the algorithms' initialization times, which were the prefetching of the first batch and the initialization of the components, respectively.

### 3.2.12 Quantification and Statistical Analysis

Statistical parameters including the definitions and exact values of  $n$  (number of frames, number of videos, or number of neurons), location and deviation measures are reported in the Figure Legends and corresponding sections in the main text. All data were expressed as mean  $\pm$  standard deviation. We used two-sided Z-test for the statistical analysis of calcium transients'  $d'$  compared to the distribution of  $d'$  values from the baseline due to noise. For all other statistical tests, we performed two-sided Wilcoxon rank sum test; n.s.: not significant, \*:  $p$ -value  $< 0.05$ , and \*\*:  $p$ -value  $< 0.005$ . We determined results to be statistically significant when  $p$ -value  $< 0.05$ . We did not remove any data from statistical analyses as outliers.

### 3.2.13 Hardware

We ran CaImAn, Suite2p, HNCcorr, and the pre- and post-processing part of our algorithm on a Windows 10 computer with Intel Xeon E5-2680 v4 CPU and 256 GB RAM. We trained and tested STNeuroNet and UNet2DS using a single NVIDIA GeForce GTX Titan X GPU. All CNNs in the CaImAn package were deployed on the NVIDIA GeForce GTX Titan X GPU.

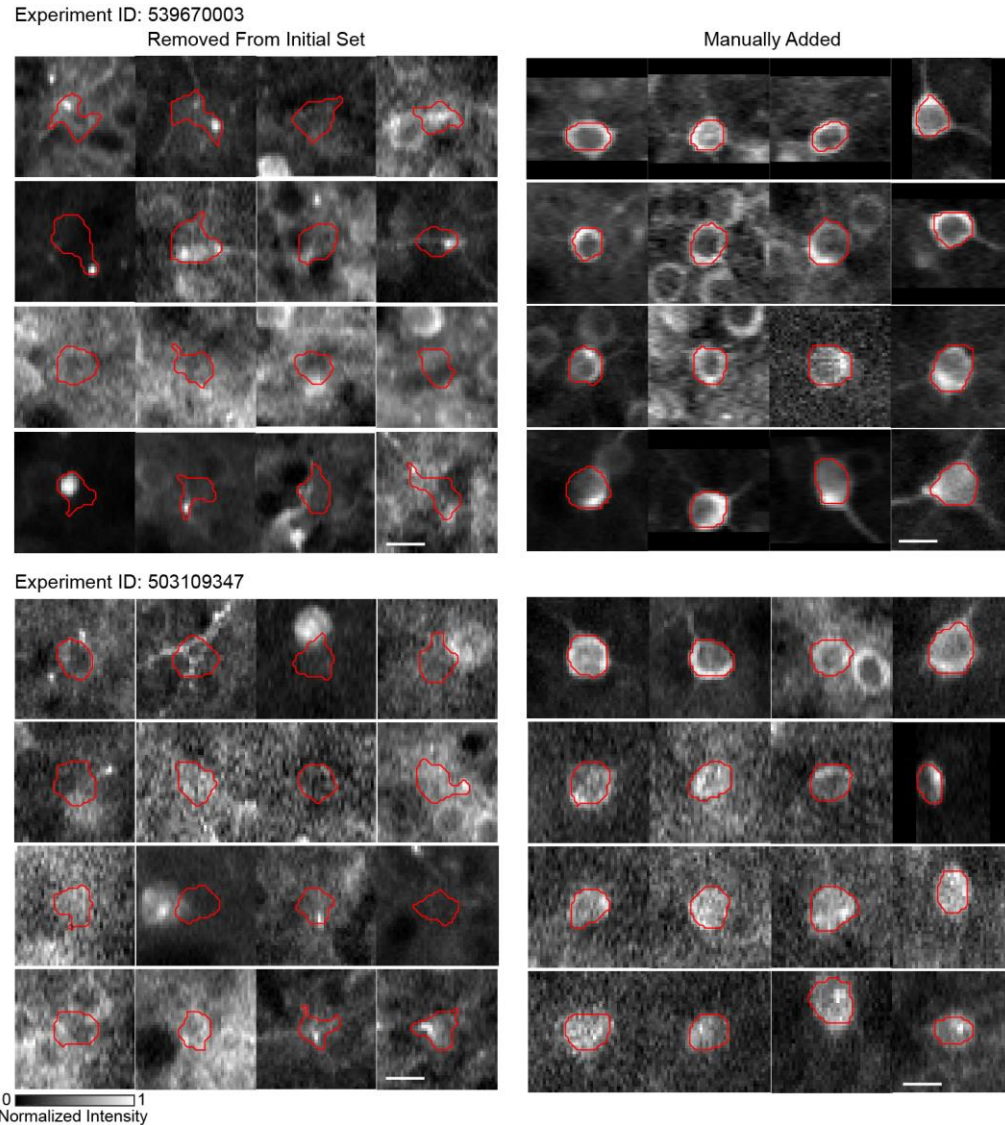
## 3.3 Results

### 3.3.1 STNeuroNet Accurately Segmented Neurons from the Allen Brain Observatory

We first quantified the performance of our method using the 275  $\mu\text{m}$  deep recordings from the ABO dataset. Upon our inspection of the provided set of masks, we

found that some of the masks did not correspond to active neurons in the selected ~13 minutes time-interval, and some active neurons were not included in the set (Figure 12). Thus, the gold-standard ground truth (GT) labels were created by our two expert human graders who sequentially edited the initial labeling. Overall, we removed  $n = 40 \pm 23.6$  masks (mean  $\pm$  standard deviation over  $n = 10$  videos) from the initial ABO marking as they were not located on the soma of active neurons, accounting for  $13.9 \pm 5.7\%$  of the initial neurons, and added  $n = 72.7 \pm 20.9$  neurons, accounting for  $24.2 \pm 5.9\%$  of the final GT neurons. The final set of neurons comprising the GT demonstrated peak calcium responses with  $d' \geq 4$  within the spike detection formalism, which were at significantly higher levels compared to the distribution of  $d'$  values from the baseline due to noise ( $p$ -values  $< 0.001$ , one-sided Z-test using  $n = 500$  baseline samples for each of the 3016 GT neurons).

Training our network on  $144 \times 144 \times 120$  segments of input data took 11.5 hours for 36,000 iterations. After training, STNeuroNet generated neuron predictions in  $171.24 \pm 21.28$  seconds (mean  $\pm$  standard deviation over  $n = 10$  videos) when processing  $4624 \pm 5$  frames of size  $487 \times 487$  pixels. The complete framework, from preprocessing to the final neuron aggregation, processed these recordings with  $17.3 \pm 1.2$  frames/s (mean  $\pm$  standard deviation over  $n = 10$  videos) speed. Note that considering the binning of videos from 30 Hz to 6 Hz, the effective processing rate can be up to 5 times better than the reported number.

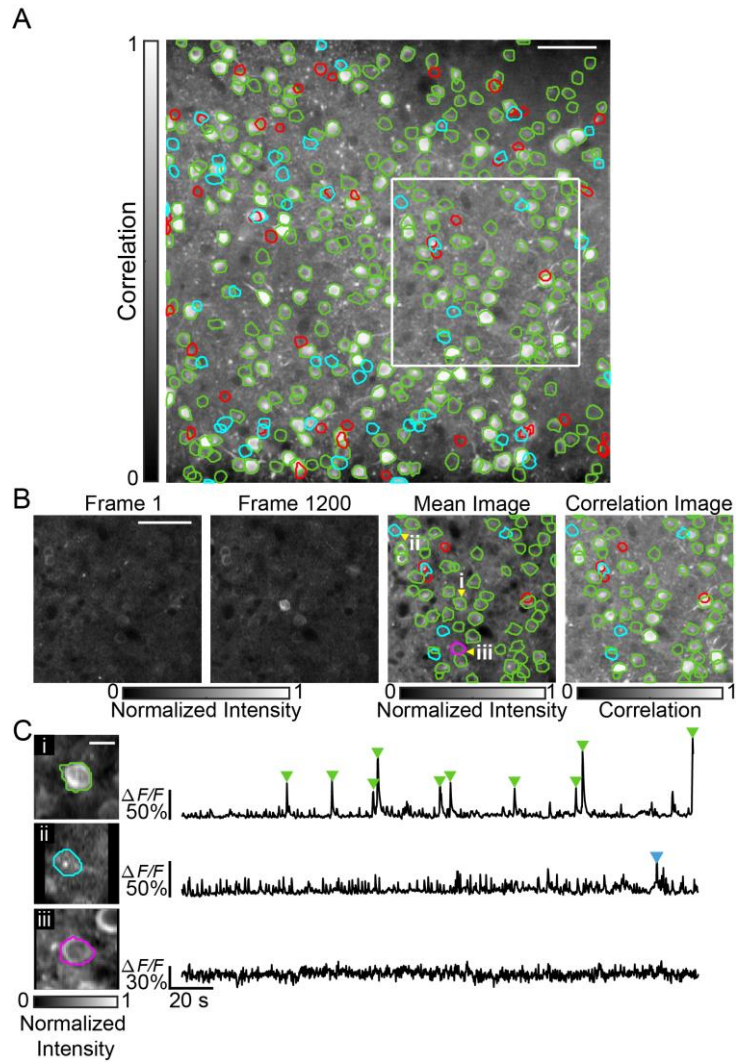


**Figure 12: Representative examples of active neuron labeling errors in the ABO dataset. Example cases that were (*left*) removed from the initial set of marking accompanied with the ABO dataset and (*right*) manually added by graders to produce the final ground truth set. Images are normalized mean of video frames at peak signal time-points. Scale bars: 10  $\mu\text{m}$ . Figure taken from [59].**

Figure 13 shows an illustrative example of our framework applied on a time-interval of  $N = 1200$  background compensated frames from one mouse, which achieved neuron detection scores (recall, precision,  $F_1$ ) of (0.86, 0.88, 0.87). The first frame, last frame, and the normalized temporal average of all frames in the batch are shown in Figure

13B. To better illustrate temporal neuronal activity, we also show the correlation image, defined as the mean correlation value between each pixel with its 4-connected neighborhood pixels. Temporal  $\Delta F/F$  traces for selected true positive, false negative, and silent neurons highlight the presence or absence of activity in the selected time-interval, indicating that STNeuroNet effectively selected active neurons while disregarding silent neurons (Figure 13B-C).

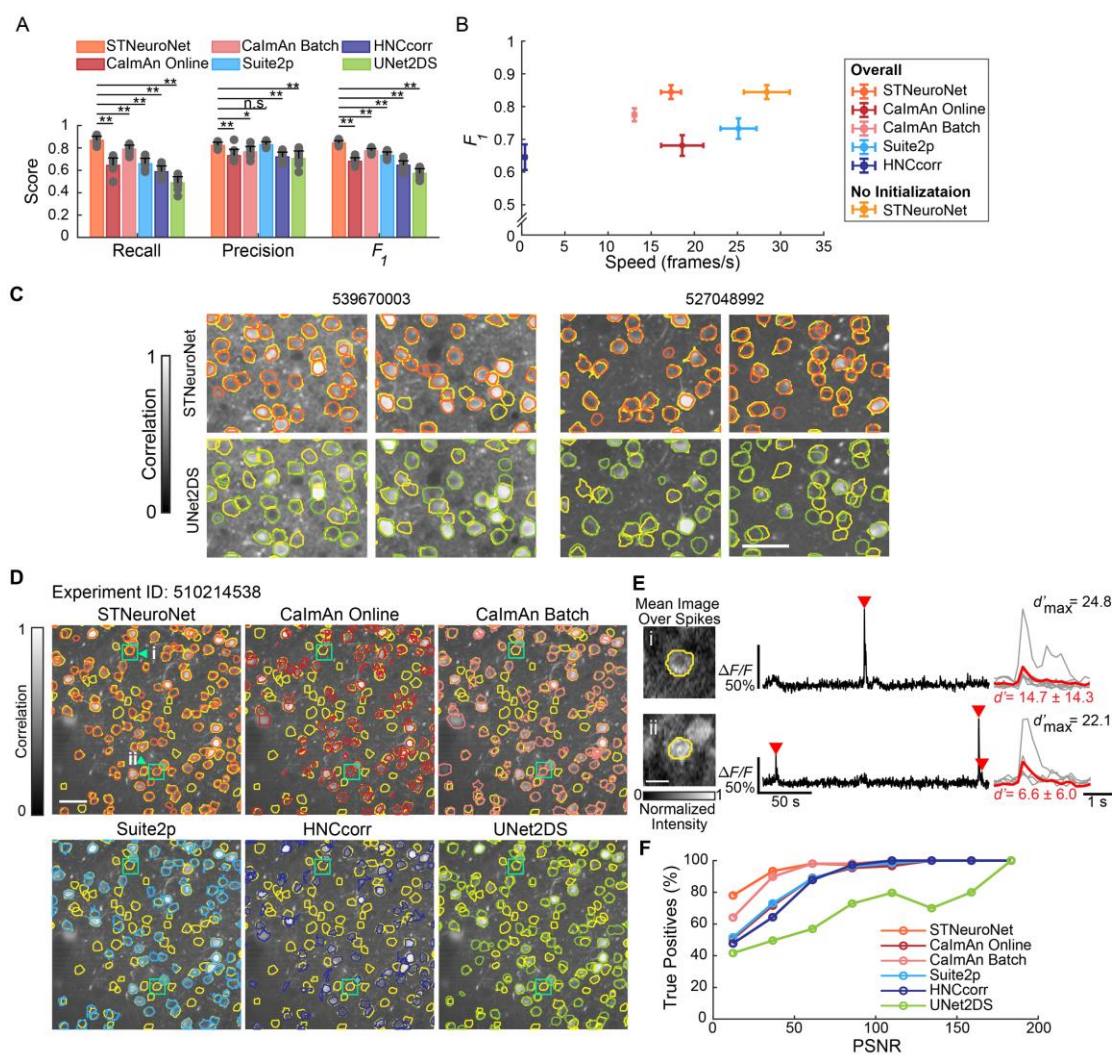
Using the same ten videos, we compared the performance of our framework to the performance of CaImAn Online and CaImAn Batch [18], Suite2p [20], HNCcorr [64], and to the deep-learning based UNet2DS [67] algorithm, quantifying each algorithm in terms of recall, precision, and  $F_1$  (Figure 14). To compare all algorithms on an equal footing, we optimized the algorithmic parameters for each method through leave-one-out cross-validation (Appendix B). Since  $F_1$  quantifies a balance between recall and precision, we used this score as the final metric to optimize and assess the performance of all methods. Our framework outperformed all other algorithms in the  $F_1$  score ( $p$ -value  $< 0.005$ , two-sided Wilcoxon rank sum test over  $n = 10$  videos; Figure 14A and Table 5) at higher speed compared to CaImAn Batch and HNCcorr ( $p$ -values  $< 0.005$ , two-sided Wilcoxon rank sum test over  $n = 10$  videos), while being as fast as CaImAn Online and slower than Suite2p ( $p$ -values = 0.3075 and  $< 0.005$ , respectively; two-sided Wilcoxon rank sum test over  $n = 10$  videos; Figure 14B) when processing  $487 \times 487$  pixels videos. After disregarding the initialization time of STNeuroNet, our framework was significantly faster than Suite2p



**Figure 13: STNeuroNet accurately identified active neurons from the ABO dataset. (A) Detection results from 1200 frames (200 seconds) of a test video overlaid on the  $200 \times 200$  pixels ( $156 \mu\text{m} \times 156 \mu\text{m}$ ) cropped region from the correlation image. Green: true positives, cyan: false negatives, and red: false positives. (B) First and last frames, normalized mean image, and correlation image from the region enclosed in the white box in A. While many neurons are visible in the mean image, only active neurons were segmented (green outlines). The neuron marked with magenta is an example silent neuron that STNeuroNet effectively disregarded. (C) Example mean images of true positive, false negative, and silent neurons (green, cyan, and magenta outlines, respectively; left) and their time-series (right) from B. Scale bars: (A-B)  $50 \mu\text{m}$  and (C)  $10 \mu\text{m}$ . Figure taken from [59].**

( $p$ -values = 0.026, two-sided Wilcoxon rank sum test over  $n = 10$  videos). For CalmAn Online, the initialization time was  $10.4 \pm 0.8$  s for 100 frames and did not contribute significantly to the total processing time. Because UNet2DS processed a single 2D image, it was extremely fast (speed =  $2263.3 \pm 2.6$  frames/s for  $n = 10$  videos), but it was not able to separate overlapping neurons, resulting in low recall values compared to other methods (Figure 14C).

We further investigated the underlying source for our framework's superior recall compared to other spatio-temporal methods. Figure 14D-E illustrate examples of sparsely-firing neurons with low  $\Delta F/F$  value calcium transients that were identified by STNeuroNet and missed by other algorithms. We further validated this observation by quantifying the percentage of GT neurons detected at different levels of PSNR in Figure 14F. STNeuroNet's higher percentage of true positive neurons compared to other algorithms in the low PSNR regime indicates that our network achieved high recall because it identified a larger portion of spiking neurons with relatively low PSNR calcium transients. On average, our algorithm detected  $22.4 \pm 7.5\%$ ,  $7.9 \pm 3.6\%$ ,  $21.0 \pm 4.8\%$ ,  $26.1 \pm 4.6\%$ , and  $38.1 \pm 5.9\%$  more neurons (mean  $\pm$  standard deviation for  $n = 10$  videos) from the GT compared to CalmAn Online, CalmAn Batch, Suite2p, HNCcorr, and UNet2DS, respectively.



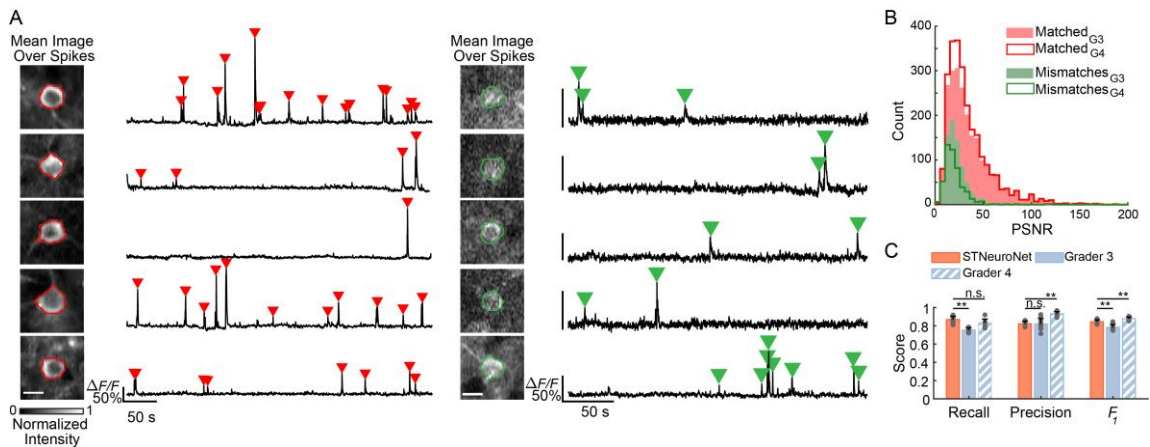
**Figure 14: STNeuroNet outperformed other methods on the ABO dataset. (A)** STNeuroNet’s performance score was significantly higher (\*:  $p$ -value < 0.05 and \*\*:  $p$ -value < 0.005, two-sided Wilcoxon rank sum test,  $n = 10$  videos). **(B)** We achieved superior detection performance at practically high processing speed. Error bars in **A** and **B** are stds for  $n = 10$  videos. **(C)** UNet2DS cannot separate overlapping neurons. **(D)** Example results of all methods. Each method’s segmented neurons are marked with different colors on top of the correlation image. Yellow markings denote the GT. **(E)** Example neurons from **D** identified by STNeuroNet and missed by other methods along with their time-series (*black traces*) and aligned activity-evoked signals (*gray traces*; *red traces*: average of gray traces). Traces are from a portion of the entire recording, with red markers denoting the times of putative calcium transients. **(F)** Percentage of detected GT neurons versus binarized PSNR. Scale bars: **(C-D)** 50  $\mu$ m and **(E)** 10  $\mu$ m. Figure taken from [59].

**Table 5: Summary of performances on all datasets. Reported numbers are in  $F_1$  (Recall, Precision) format, where in each field we report the mean  $\pm$  standard deviation across  $n = 10$  and  $n = 6$  videos for the ABO and Neurofinder dataset, respectively.**

	<i>ABO Layer 275</i> <i>(cross-validation)</i>	<i>ABO Layer 175</i>	<i>Neurofinder Test</i>
<i>STNeuroNet</i>	0.84 $\pm$ 0.02 (0.87 $\pm$ 0.04, 0.82 $\pm$ 0.03)	0.86 $\pm$ 0.03 (0.86 $\pm$ 0.03, 0.85 $\pm$ 0.04)	0.70 $\pm$ 0.03 (0.82 $\pm$ 0.07, 0.61 $\pm$ 0.03)
<i>CaImAn</i> <i>Online</i>	0.68 $\pm$ 0.03 (0.64 $\pm$ 0.07, 0.73 $\pm$ 0.06)	0.62 $\pm$ 0.05 (0.55 $\pm$ 0.07, 0.72 $\pm$ 0.05)	0.53 $\pm$ 0.09 (0.50 $\pm$ 0.10, 0.58 $\pm$ 0.10)
<i>CaImAn</i> <i>Batch</i>	0.77 $\pm$ 0.02 (0.79 $\pm$ 0.04, 0.76 $\pm$ 0.05)	0.75 $\pm$ 0.03 (0.75 $\pm$ 0.05, 0.74 $\pm$ 0.03)	0.62 $\pm$ 0.05 (0.67 $\pm$ 0.06, 0.58 $\pm$ 0.06)
<i>Suite2p</i>	0.73 $\pm$ 0.03 (0.66 $\pm$ 0.05, 0.83 $\pm$ 0.03)	0.67 $\pm$ 0.08 (0.62 $\pm$ 0.09, 0.73 $\pm$ 0.08)	0.61 $\pm$ 0.08 (0.64 $\pm$ 0.12, 0.61 $\pm$ 0.15)
<i>HNCcorr</i>	0.65 $\pm$ 0.04 (0.59 $\pm$ 0.05, 0.72 $\pm$ 0.04)	0.59 $\pm$ 0.06 (0.58 $\pm$ 0.08, 0.62 $\pm$ 0.07)	0.47 $\pm$ 0.08 (0.43 $\pm$ 0.07, 0.53 $\pm$ 0.12)
<i>UNet2DS</i>	0.57 $\pm$ 0.04 (0.49 $\pm$ 0.06, 0.71 $\pm$ 0.07)	0.59 $\pm$ 0.04 (0.55 $\pm$ 0.04, 0.65 $\pm$ 0.07)	0.49 $\pm$ 0.10 (0.46 $\pm$ 0.15, 0.58 $\pm$ 0.14)

To assess the reproducibility of our GT markings, we trained a third grader to conduct an inter-human agreement test. Grader #3 labelled these data from scratch without access to the initial masks from the Allen Institute or the consensus GT segmentations produced by the first two graders. GT and grader #3 were consistent in segmenting neurons with high PSNR (Figure 15A-B). The resulting distribution of mismatched cases (set of missed and falsely-labelled neurons) was weighted towards neurons with low PSNR values, which challenge human perception during manual marking of the video (Figure 15B). Our framework achieved a higher  $F_1$  score compared to grader #3 (mean of 0.84 vs 0.78,  $p$ -value = 0.0013; two-sided Wilcoxon rank sum test for  $n = 10$  videos; Figure 15C). To mimic the case of semi-automatic marking, we asked a

fourth grader to independently correct the ABO markings for these videos. Compared to grader #4, both grader #3 and STNeuroNet achieved lower F<sub>1</sub> scores ( $p$ -values = 0.0002 and 0.0036, respectively; two-sided Wilcoxon rank sum test for  $n = 10$  videos; Figure 15C), which is due to the inherent bias between the GT set and grader #4's markings.



**Figure 15: Inter-human agreement test for ABO neuron segmentation.** (A) Examples of common neurons between GT and grader #3 (red data) and missed neurons by grader #3 (green data). Normalized mean images of neurons over their active time-intervals are shown. Missed neurons exhibit low peak  $\Delta F/F$  or atypical appearance. Scale bars: 10  $\mu\text{m}$ . (B) Histogram of the PSNR for mismatched (green data) and matched neurons (red data) between GT and graders #3 and #4. (C) Our algorithm achieved similar recall score as grader #4 ( $p$ -value = 0.0757) and similar precision as grader #3 ( $p$ -value = 0.6776) resulting in a F<sub>1</sub> score between that of the two graders (n.s.: not significant; \*\*:  $p$ -value < 0.005). Figure taken from [59].

### 3.3.2 Trained STNeuroNet Segmented Neurons from Unseen Recordings of Additional Cortical Layers

To demonstrate the generalizability of our trained STNeuroNet, we next applied our segmentation framework to recordings from a different cortical layer in VISp. We trained STNeuroNet with the same ten videos as in the previous section, from 275  $\mu\text{m}$

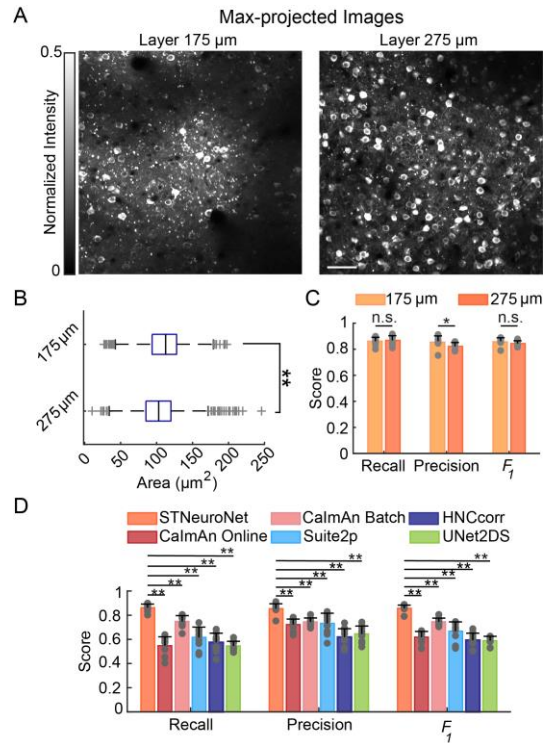
below the pia in VISp. The neurons in these datasets were drawn from the Rorb-IRES2-Cre mouse line, which restricts expression to layer 4 neurons, and the Cux2-CreERT2 mouse line, which restricts expression to excitatory cell types. We tested this network on data acquired from ten different mice, this time from a different cortical layer at 175  $\mu\text{m}$  deep in VISp. The neurons in these datasets were drawn from the Cux2-CreERT2 and Emx1-IRES-Cre mouse lines, which express calcium sensors in excitatory neurons (Table 4).

The data from 175  $\mu\text{m}$  deep is putatively in layer 2/3, while the data from 275  $\mu\text{m}$  deep is at the interface between layer 4 and layer 2/3. Neurons from the test dataset were qualitatively visually different from neurons in the training set (Figure 16A). Quantitatively, the test set had bigger neurons (median of 112.6  $\mu\text{m}^2$  versus 102.8  $\mu\text{m}^2$ ;  $p$ -value  $< 0.005$ , two-sided Wilcoxon rank sum test over  $n = 2182$  and 3016 neurons, respectively; Figure 16B) and lower densities of identified active neurons ( $0.0014 \pm 0.0002$  neurons/ $\mu\text{m}^2$  versus  $0.0019 \pm 0.0003$  neurons/ $\mu\text{m}^2$  for 175 and 275  $\mu\text{m}$  data, respectively;  $p$ -value  $< 0.005$ , two-sided Wilcoxon rank sum test over  $n = 10$  videos). Despite the differences in the size and density of neurons within these two datasets, our network trained on 275  $\mu\text{m}$  data performed at indistinguishable levels on 275  $\mu\text{m}$  test data and 175  $\mu\text{m}$  data ( $p$ -value = 0.1212 for  $F_1$ ; two-sided Wilcoxon rank sum test with  $n = 10$  videos for both groups; Figure 16C and Table 5). Using the layer 275  $\mu\text{m}$  data to set the algorithmic parameters of other methods, our framework achieved the highest mean  $F_1$  score on the 175  $\mu\text{m}$  data ( $p$ -value  $< 0.005$ , two-sided Wilcoxon rank sum test over  $n = 10$  videos; Figure

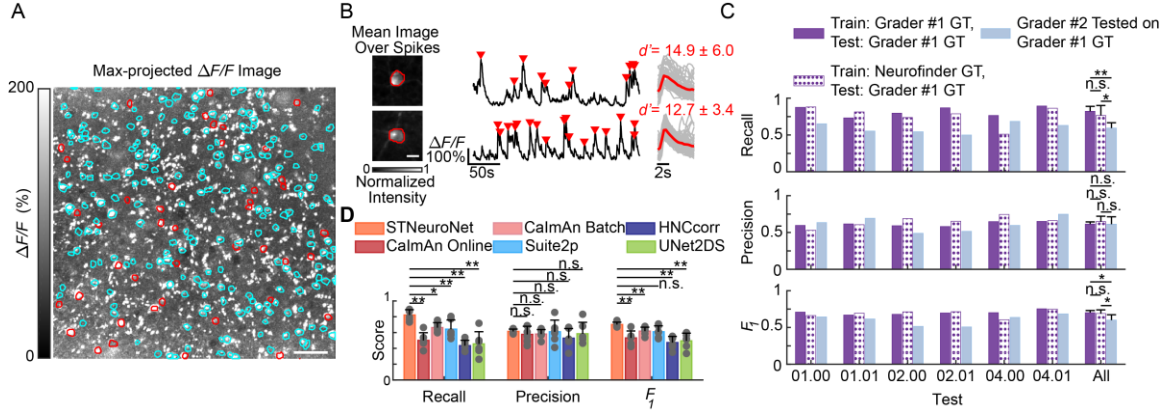
16D). Unlike our method, the  $F_1$  scores of all other methods except UNet2DS were significantly lower on the 175  $\mu\text{m}$  data compared to the 275  $\mu\text{m}$  test data ( $p$ -values = 0.006, 0.031, 0.021, and 0.045 for CaImAn Online, CaImAn Batch, Suite2p, and HNCcorr, respectively; two-sided Wilcoxon rank sum test over  $n = 10$  videos; Table 5).

### 3.3.3 STNeuroNet Accurately Segmented Neurons from the Neurofinder Dataset

We also applied our framework on two-photon calcium imaging data from the Neurofinder challenge. We used this dataset with activity-informed markings for training and comparison between different algorithms. Upon systematic inspection of Neurofinder GT sets, we found many putative neurons ( $n = 2, 2, 81, 60, 50$  and 19 neurons for datasets called 01.00, 01.01, 02.00, 02.01, 04.00, and 04.01, respectively, corresponding to 0.5%, 0.6%, 41.1%, 33.7%, 21.2%, and 7.67% of the original GT neurons) with spatial shape and fluorescence temporal waveforms expected from GCaMP6-expressing neurons. Examples of such GT errors from the data called 04.00 in the Neurofinder training set are shown in Figure 17A-B. The extracted transients in the time-series of newly-found neurons among all datasets had high detectability index  $d' > 3.2$ , emphasizing that these signals are truly activity-evoked transients. We also computed the average fluorescence image during these highly detectable transients, which yielded high quality images of the neurons (Figure 17B left).



**Figure 16: Trained STNeuroNet performed equally well on data from a different cortical layer and outperformed other methods. (A) Qualitative comparison between Layer 275  $\mu\text{m}$  and 175  $\mu\text{m}$  data. Images are the normalized maximum-projection images over the entire recording. (B) The area of neurons labeled from the two cortical depths were different (\*\*:  $p$ -value  $< 0.005$ ;  $n = 2182$  and  $3016$  neurons from the 175  $\mu\text{m}$  and 275  $\mu\text{m}$  datasets, respectively), with the higher depth exhibiting smaller neurons. (C) The performance scores were not significantly different for recall and  $F_1$  ( $p$ -values =  $0.5708$  and  $0.1212$ , respectively; \*:  $p$ -value  $< 0.05$ ) between the two datasets using the network trained on the 275  $\mu\text{m}$  data. (D) STNeuroNet's performance on the 175  $\mu\text{m}$  data was superior compared to other methods (\*:  $p$ -value  $< 0.05$  and \*\*:  $p$ -value  $< 0.005$ ). Figure taken from [59].**



**Figure 17: STNeuroNet achieved best performance in the Neurofinder challenge, which contained suboptimal markings. (A) Overlay of the initial GT neurons (cyan outlines) and the added neurons after manual inspection (*red outlines*) on the maximum-projection of  $\Delta F/F$  from the 04.00 training data. (B) Example neurons missed by the Neurofinder-supplied markings from A along with their neuropil-subtracted time-series (*black traces*). These neurons exhibit transients (*gray traces*: temporally aligned activity-evoked signals; *red traces*: average of gray traces) typical of GCaMP6-expressing neurons with high detection fidelity  $d'$ . Images are the average of frames within active time-intervals. (C) When tested on grader #1's GT, STNeuroNet's performance was not significantly different from when it was trained on either of Neurofinder's GT or grader #1's GT ( $p$ -value = 0.9372). Both networks achieved above-human performance in average  $F_1$  score across the test dataset compared to grader #2, when tested with grader #1's GT ( $p$ -values = 0.0411 and 0.0087). (D) STNeuroNet outperformed other methods as denoted by the average  $F_1$  score (\*:  $p$ -value < 0.05 and \*\*:  $p$ -value < 0.005). Scale bars: (A) 50  $\mu$ m and (B) 10  $\mu$ m. Figure taken from [59].**

We analyzed the impact of using different training GT sets on STNeuroNet's performance. The senior grader (grader #1) corrected the labeling of the training data by adding the missing neurons to the GT sets and labelled the Neurofinder test set. Compared to the case of using Neurofinder's GT for training, the average  $F_1$  score was not significantly different to the case of employing the markings from grader #1 for both

training and testing ( $p$ -value = 0.9372, two-sided Wilcoxon rank sum test over  $n = 6$  videos; Figure 17C). Similar to the ABO dataset, we conducted an inter-human agreement test. Independent from grader #1, grader #2 created a second set of markings for the test datasets. When tested on grader #1's markings, our algorithm attained comparable average  $F_1$  score to grader #2 ( $p$ -value = 0.2403 and 0.3095 for training on Neurofinder's GT and grader #1's GT, respectively; two-sided Wilcoxon rank sum test over  $n = 6$  videos; Figure 17C).

Using our expert manual markings as GT for the Neurofinder dataset, we compared our framework to other methods (Figure 17D and Table 5). For all algorithms, we used the entire Neurofinder training set to optimize the algorithmic parameters for each method. Our framework (STNeuroNet trained with the entire training set) achieved higher but statistically insignificant  $F_1$  score than Suite2p (mean  $\pm$  standard deviation of  $0.70 \pm 0.03$  and  $0.61 \pm 0.08$ , respectively;  $p$ -value = 0.0649, two-sided Wilcoxon rank sum test over  $n = 6$  videos). Compared to all other methods, STNeuroNet's  $F_1$  score was significantly higher ( $p$ -values  $< 0.005$ , two-sided Wilcoxon rank sum test over  $n = 6$  videos).

To further test the generalizability of our framework to data acquired with different experimental conditions, we compared the performance of STNeuroNet trained on the ABO Layer 275  $\mu\text{m}$  dataset to STNeuroNet trained on all Neurofinder training set, when evaluated on the Neurofinder test data (Table 6). Although using the ABO Layer 275  $\mu\text{m}$  data for training resulted in lower mean  $F_1$  score, the scores were not statistically different ( $p$ -value = 0.485, two-sided Wilcoxon rank sum test for  $n = 6$  videos), and the

performance was comparable to that of Suite2p ( $p$ -value = 1, two-sided Wilcoxon rank sum test for  $n = 6$  videos). With the addition of the high-quality ABO Layer 275  $\mu\text{m}$  data to the Neurofinder training set, STNeuroNet achieved higher  $F_1$  score compared to the network trained only on the Neurofinder training set ( $p$ -value = 0.026, two-sided Wilcoxon rank sum test for  $n = 6$  videos; Table 6).

**Table 6: STNeuroNet performance on all data when trained on different datasets. Reported numbers are in  $F_1$  (Recall, Precision) format, where in each field we report the mean  $\pm$  standard deviation across  $n = 10$  and  $n = 6$  videos for the ABO and Neurofinder datasets, respectively.**

Test Data	Train Data			
	<i>ABO Layer 275</i>	<i>Neurofinder Train</i>	<i>ABO Layer 275 + Neurofinder Train</i>	<i>All ABO</i>
<i>ABO Layer 275</i>	0.84 $\pm$ 0.02 <sup>a</sup> (0.87 $\pm$ 0.04, 0.82 $\pm$ 0.03)	0.74 $\pm$ 0.04 (0.86 $\pm$ 0.02, 0.64 $\pm$ 0.05)	N/A	N/A
<i>ABO Layer 175</i>	0.86 $\pm$ 0.03 (0.86 $\pm$ 0.03, 0.85 $\pm$ 0.04)	0.74 $\pm$ 0.04 (0.82 $\pm$ 0.05, 0.68 $\pm$ 0.05)	0.85 $\pm$ 0.03 (0.88 $\pm$ 0.03, 0.82 $\pm$ 0.05)	N/A
<i>Neurofinder Test</i>	0.62 $\pm$ 0.17 (0.52 $\pm$ 0.24, 0.88 $\pm$ 0.08)	0.70 $\pm$ 0.03 (0.82 $\pm$ 0.07, 0.61 $\pm$ 0.03)	0.75 $\pm$ 0.04 (0.72 $\pm$ 0.11, 0.79 $\pm$ 0.04)	0.67 $\pm$ 0.11 (0.60 $\pm$ 0.21, 0.83 $\pm$ 0.09)
<i>Neurofinder Train</i>	0.48 $\pm$ 0.13 (0.37 $\pm$ 0.18, 0.83 $\pm$ 0.13)	N/A	N/A	0.55 $\pm$ 0.11 (0.45 $\pm$ 0.18, 0.80 $\pm$ 0.13)

<sup>a</sup>: Performance quantified with leave-one-out cross-validation.

### 3.4 Discussion

We presented an automated, fast, and reliable active neuron segmentation method to overcome a critical bottleneck in the analysis workflow of utilizing neuronal signals in real-time behavioral studies. The core component of our method was an efficient 3D CNN named STNeuroNet. The performance of this core was further improved by intuitive pre- and post-processing steps. Our proposed framework for sequential processing of the

entire video accurately segmented overlapping active neurons. In the ABO dataset, our method surpassed the performance of CaImAn, Suite2p, HNCcorr, UNet2DS, and an expert grader, and generalized to segmenting active neurons from different cortical layers and regions with different experimental setups. We also achieved the highest mean  $F_1$  score on the diverse datasets from the Neurofinder challenge.

STNeuroNet is an extension of DenseVNet [78], which consists of 3D convolutional layers, to segment active neurons from two-photon calcium imaging data. The added temporal max-pooling layer to the output of DenseVNet summarized the spatio-temporal features into spatial features. This step greatly increased the speed of training and inference processes, which is important for high-speed network validation and low-latency inference in time-sensitive applications such as closed-loop experiments.

We showed the superior performance of our method for active neuron detection and segmentation by direct comparison to the state-of-the-art classic machine learning as well as deep learning methods. We achieved this level of performance by consistently detecting larger number of true active neurons compared to other algorithms. Our superior performance was not dependent on the GT created by graders #1 and #2. This is in part due to the fact that unlike the model-based spatiotemporal deconvolution methods of CaImAn and Suite2p, our proposed STNeuroNet extracts relevant spatiotemporal features from the imaging data without prior modeling; the deep-learning approach could be more flexible for detecting arbitrary spatiotemporal features. Compared to the deep learning based UNet2DS that is applied to a single aggregate (mean) image, our proposed

framework was more powerful in discriminating overlapping neurons and identifying neurons with low activity-evoked contrast because it assesses information in each video frame individually, and in concert with other frames.

One advantage of deep learning-based methods is that once trained, they are computationally fast at inference time. We showed that our framework achieved significantly higher detection scores compared to all other methods at practically high processing speed. While we measured the computational speed of all algorithms on the same computer, we acknowledge that some of these algorithms could potentially benefit from more computationally optimal coding that target other specific hardware architectures. Combined with signal separation [14, 29, 85] and fast spike detection algorithms [4, 29, 35], our framework could potentially enable fast and accurate assessment of neural activity from two-photon calcium imaging data. Our current implementation performed neuron detection at near video-rate processing of individual frames when processing sets of sequential frames, which suggests that our framework can interleave updates of the segmentation results with data acquisition. Because our framework can be applied to overlapping or non-overlapping temporal batches, it presents a flexible trade-off to either increase speed or accuracy: processing non-overlapping temporal batches speeds up the algorithm, while using the median or mean probability map of highly overlapping batches could potentially improve the performance at inference time.

Depending on the complexity of the problem and the architecture of neural networks, deep-learning methods need different amount of training data to achieve high performance scores and to be generalizable. We utilized data augmentation, dropout [80], and batch-normalization [79] to achieve generalizability and prevent overfitting. We demonstrated the generalizability of our trained STNeuroNet by applying the processing framework on recordings from different cortical layers and regions. We were able to train STNeuroNet on neurons from 275  $\mu\text{m}$  deep in the mouse cortex and segment active neurons from 175  $\mu\text{m}$  deep at an indistinguishable performance level, despite the differences in the neuron size and densities at these two depths. This experiment confirmed that our network was not over-trained to segment active neurons from a specific cortical depth. Adding ABO Layer 275  $\mu\text{m}$  data to the Neurofinder training dataset improved accuracy of segmenting the Neurofinder test dataset. These results suggest that utilizing training data acquired with different experimental setups is beneficial for generalizing STNeuroNet. Also, training on the entire ABO dataset and testing on Neurofinder recordings shows that having more training data from one experimental set up improves performance of segmenting videos from a different experimental set up. These experiments confirm that other neuroscientists with significantly different recordings can take advantage of our trained network through transfer learning [86] to adapt the network to their specific data. Combined with transfer learning, our trained network has the potential to achieve high performance and generalizability on experimentally diverse recordings.

In this work, we carefully relabeled active neurons from the ABO dataset to compare the performance of different algorithms. To minimize the probability of human error in marking active neurons, we created the final set of GT masks by combining the markings from two independent graders. To assess human grading consistency, we compared the markings of a third independent grader performing manual segmentation from scratch to the GT. We showed that our framework’s performance was higher than grader #3’s, suggesting that STNeuroNet learned informative features and surpassed human-level accuracy in active neuron segmentation. For the sake of completeness, we added an additional experiment to reflect the effect of bias in performance of human graders. We compared our method to grader #4, a grader who corrected the ABO dataset markings with similar procedures to, but independently of, graders #1 and #2. As expected, due to the bias created by having access to pilot segmentation labels, grader #4’s markings were closer to the GT than grader #3’s markings.

Naturally, using manual labeling as the gold-standard has the disadvantage of introducing human errors and bias in the GT data. However, currently available alternative approaches are even less suitable for generating GT. For example, simultaneous dual channel imaging of activity-independent nuclear tagged neurons provides reliable ground truth markings for all neurons. However, such labels which include both active and inactive neurons are not suitable for evaluating segmentation methods for active neurons in behavioral experimentations. Progress in activity-based

neuron labeling methods combined with simultaneous optical and structural imaging techniques may provide reliable gold-standard datasets in future.

In addition to the ABO dataset, we also included the results of segmenting the diverse Neurofinder challenge datasets. We included these results because the Neurofinder dataset has been used to assess the accuracy of many recent segmentation algorithms [20, 63, 64, 67]. Our framework significantly outperformed all other methods except Suite2p, which could be due to the small sample size and the relatively large spread of Suite2p’s  $F_1$  scores. It is encouraging that our method achieved the highest mean  $F_1$ , but our finding that the GT labeling of the training dataset from the challenge has missed neurons is nearly as important. While we do not have access to the labeling of the test dataset, we presume that GT accuracies in the publicly available training datasets match that of the test data. Thus, we carefully manually labeled the test set in the Neurofinder challenge. The availability of these carefully labeled GT training and test sets are expected to improve the fairness and accuracy of the evaluation metrics to be used for assessing future segmentation algorithms. Similar to the ABO dataset, we achieved above-human-level performance when training on our carefully labeled markings. Furthermore, when using our carefully curated test labels to evaluate the performance of STNeuroNet under different training conditions, we found that training on our carefully curated training labels only marginally improved performance when compared to training on Neurofinder’s labels. This might be due to the nature of the CNN architecture. The architecture seeks to establish a complex yet consistent pattern in data and could average

out erroneous labeling of a subset of the training set as outliers. However, errors in labeling of the test set more affect the performance metrics, as experimentalists use these erroneous labels to directly evaluate the network's output. The impact of training with noisy or incorrect labels on the performance of CNNs is still the subject of active research [87-89], and an in-depth analysis of their effect was beyond the scope of this work.

We also note that regardless of correct labeling, the limited number of training samples per dataset in the Neurofinder challenge is a major bottleneck for optimal training of CNN-based methods. Our method achieved generalizability and human-level performance, and thus, could assist in the creation of additional accurate training sets for future algorithm development. CNN-generated GT datasets could potentially reduce the workload of human graders while improving the accuracy of the markings by minimizing human errors due to subjective heuristics.

This work is the first step in a continuum of research to utilize 3D CNNs for detection and segmentation of neurons from calcium imaging data. The data used in our work were properly corrected for motion artifacts by the data owners. In the more general case of non-registered datasets, algorithms such as NoRMCorre [90] can be used to accurately correct motion prior to the application of our framework. We used watershed to separate the identified overlapping neurons co-activated in the same time interval processed by STNeuroNet, which can give inaccurate masks. Since such overlapping neurons might segregate themselves in other time intervals, we presented the neuron fusion process to circumvent this issue and obtain masks that had overlapping pixels.

Each component of our method, individually or together, can be used by us and other researchers in many related projects. To this end, as our computationally fast and accurate method is an invaluable tool for a large spectrum of real-time optogenetic experiments, we have made our open-source software and carefully annotated datasets freely available online. Future work will extend the current framework to minimize parameter adjustments in pre- and post-processing steps by encapsulating these steps into an end-to-end learning process. Such an approach would remove the need for watershed-based separation of overlapping neurons, which is prone to error for one-photon recordings or two-photon imaging of species or brain areas with significantly overlapping populations, which was not present in the data utilized in our work.

## 4. Instance Segmentation of Ganglion Cells from AO-OCT Volumes via Weakly-Supervised Deep Learning

The purpose of this chapter was to develop a reliable automatic ganglion cell (GC) counting and segmentation method from adaptive optics optical coherence tomography (AO-OCT) volumes of human retina. The material in this chapter were partially presented at the *SPIE Photonics West BiOS Conference on Ophthalmic Technologies XXX* under the title “Fully automatic quantification of individual ganglion cells from AO-OCT volumes via weakly supervised learning” [91] and *ARVO annual meeting* under the title “Automatic cellular level differentiation of glaucomatous and healthy eyes via deep learning-based adaptive optics OCT analysis” [92]. The full material of this chapter has been submitted for journal publication. The contents in this chapter including texts, figures, and tables were mainly reproduced from these manuscripts.

Here we present an automated GC layer (GCL) soma quantification method from AO-OCT volumes based on weakly-supervised deep learning. We show that: (1) for localizing GCL somas from healthy and glaucoma subjects, our method performed on par or superior to human experts and (2) the measures of soma diameters were in line with previous histological and semi-automatic *in vivo* studies. These results suggest that our method should have broad appeal for long-term investigations of GC populations.

### 4.1 Introduction

Ganglion cells are one of the fundamental retinal neurons that process and transmit vision information to the brain. These cells are degenerated in optic

neuropathies, such as glaucoma, which can lead to irreversible blindness if not managed properly [22, 93]. In glaucoma clinical practice, visual function measured through standard automated perimetry and structural measurements using fundus imaging and OCT are utilized for diagnosis and monitoring of the disease. The visual field test is subjective, has poor sensitivity to early disease [22, 94], and its high variability limits reliable identification of vision loss [95-97]. OCT has been increasingly incorporated into clinical practice to improve disease care, with the thickness of the nerve fiber layer (NFL) as a widely used metric [98-100]. While the NFL is composed of GC axons, it also contains significant glial tissue, which varies across the retina [101], and at advanced stages of glaucoma, the NFL thickness reaches a nadir despite continued progression of the disease [102-104]. Alternatively, the ganglion cell complex (GCC; comprised of the NFL, GC layer, and the inner plexiform layer) thickness has been suggested as a better candidate for monitoring glaucoma progression [102]. Although the GCC thickness measured through OCT is promising, it still reflects the coarse aggregate of underlying cells, and therefore does not finely capture soma loss and morphology changes at the cellular level. Since using one of the aforementioned data alone does not provide a complete picture of glaucomatous damage, more recent studies have employed different combinations of these structural and functional datasets – some with machine learning approaches – to assess disease damage [105-108]. Study of these methods remain ongoing.

In principle, the ability to quantify features of individual GCs offers the potential of highly sensitive biomarkers for improved diagnosis and treatment monitoring of GC

loss in glaucoma and other neurodegenerative diseases. The recent incorporation of adaptive optics (AO) with OCT [24, 25, 109, 110] and scanning light ophthalmoscopy (SLO) [111] allows visualization of GC layer (GCL) somas in the living human eye. While successful, the current standard approach for quantification – manual marking of AO-OCT volumes – is subjective, time-consuming, and not practical for large-scale studies and clinical use. Thus, there is a need to develop an automatic technique for rapid, high-throughput, and objective identification of GCL somas and quantification of their morphological properties.

To date, many automated methods for localizing various retinal structures and cell types from different ophthalmic imaging systems have been proposed. Segmentation of retinal layer boundaries from OCT images [112-116], segmentation of retinal blood vessels [117-123], detection of photoreceptors from AO images [124-129], and segmentation of various anatomical and pathological features [115, 126, 130-132] are just few examples of previous work in the field of ophthalmic image processing. These methods span from mathematical model-based techniques to the more recent deep learning-based algorithms. In the field of deep learning, convolutional neural networks (CNNs) have become a staple in image analysis tasks such as classification and segmentation due to their exceptional performance. Previous deep learning-based ophthalmic image processing studies mainly used CNNs with 2-dimensional (2D) filters to segment different retinal structures. However, depending on the imaging system resolution and sampling scheme, some structures like GCs cannot be summarized into a single 2D image as they occupy a large

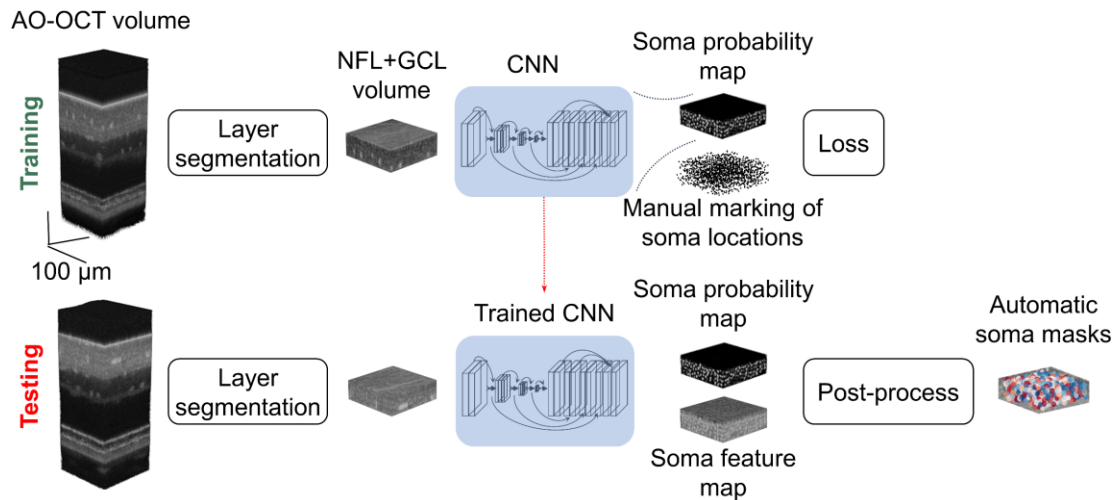
3-dimensional (3D) space. Therefore, CNNs that make use of 3D context information could be superior to 2D CNNs when processing volumetric data. To incorporate 3D information into the learning process, some studies have treated multiple adjacent slices of a volume as different channels in the input layer of their CNN. Although this approach accepts a volume as the input, the first convolutional layer collapses the 3D input into multiple 2D feature maps. An alternative for exploiting 3D contextual cues is to utilize 3D convolutional operations in the CNN. In the past few years, CNNs with 3D filters have been successfully applied to various medical image processing applications [73-75, 133-136] and spatiotemporal analysis of videos [59, 71, 72].

Fully supervised training of CNNs usually requires large training datasets to achieve acceptable performance. To address this problem for detecting photoreceptors from AO-OCT images, Heisler *et al.* [129] used transfer learning by taking advantage of existing manually-labeled AO scanning light ophthalmoscope datasets. Unfortunately, a dataset of volumetric manually segmented GCs from any imaging system does not currently exist. Adding to the difficulty of training CNNs, the pixel-level annotations needed for semantic segmentation is a strenuous task for densely packed GCs in AO-OCT volumes. Currently there is growing interest in weakly supervised segmentation schemes using different levels of weak annotation such as image-level labels [137, 138], scribbles [139-141], bounding boxes [142-145], and click-points [146]. Studies that use image-level labels for instance segmentation often utilize class activation maps [147, 148] to localize objects in the image. After localization, these methods use a segmentation proposal

technique (e.g. multi-scale combinatorial grouping [149]) to obtain the final object masks. Other group of studies often combine graphical models with bounding boxes or seeds to obtain initial object masks for fully supervised training. Although the segmentation masks are iteratively updated during the training of the segmentation network as described in [142, 143], errors in the initial training steps of this approach could negatively affect the training process [150, 151]. To avoid this problem, Tang et al. [140, 141] incorporated criteria from unsupervised segmentation techniques into the training loss function. They used partial cross-entropy loss on labeled pixels and a normalizedcut-based [140] and conditional random field (CRF)-based loss [141] for unlabeled pixels. Such intricate measures are often necessary for weakly supervised segmentation of objects with complex structures frequently present in natural images. As we show, we can obtain accurate GCL soma segmentation masks from our CNN trained on click-points using straightforward post-processing steps.

Here we present the first automatic deep learning-based method for localizing GCL somas and measuring their diameters in independent sets of AO-OCT volumes acquired with two imaging devices from healthy and glaucoma subjects. Unlike fully-supervised, multi-task CNNs, we trained our network only with click-points in the context of weakly-supervised training to segment individual GCL somas (instance segmentation of GCL somas). We demonstrate that our method achieved high detection performance on par with expert graders and diameter measurements in line with previous histological and *in vivo* semi-automatic measurement studies. In addition, our method

was generalizable to a previously unseen retinal location between both imaging devices and groups of subjects. Experiments on glaucoma subjects showed that in addition to expert-level performance, cellular-level measurements correlated with GCL thickness and can potentially aid in disease diagnosis.



**Figure 18: Overview of the weakly supervised deep learning method for GC segmentation. During training, after segmenting the GCL from the entire AO-OCT volume, the volumes along with the manually marked soma locations are fed into the CNN. The CNN outputs are then post-processed to segment individual somas. At test time, we use the trained CNN and post-processing parameters on new AO-OCT volumes.**

## 4.2 Methods

We developed and trained a weakly supervised deep learning-based framework for segmenting individual GCL somas from AO-OCT volumes (Figure 18). Briefly, we first narrowed the search space for GCL somas by automatically extracting this retinal layer from the complete AO-OCT volume (section 4.2.4.1 Data pre-processing). During the CNN training phase, instead of directly training our CNN (Figure 20, see section 4.2.4.2 Neural network and training process) to learn the segmentation task, the CNN was

trained to localize GCL somas using manually marked soma locations. At testing phases, the network’s output volumes were used to segment individual somas with additional post-processing steps (see section 4.2.4.3 Soma localization and segmentation). In the next sections, we present details about the data and each step of our framework.

#### 4.2.1 AO-OCT Datasets

We used two separate datasets acquired by the AO-OCT systems developed at Indiana University (IU) and the U.S. Food and Drug Administration (FDA), previously described[24][25]. Briefly, IU’s resolution in retinal tissue was  $2.4 \times 2.4 \times 4.7 \mu\text{m}^3$  (width  $\times$  length  $\times$  depth), with width and length specified by the Rayleigh resolution limit. The dataset consisted of  $1.5^\circ \times 1.5^\circ$  AO-OCT volumes from eight healthy subjects (seven males and one female, age:  $32.4 \pm 10.6$  years) at  $3^\circ$ - $4.5^\circ$ ,  $8^\circ$ - $9.5^\circ$ , and  $12^\circ$ - $13.5^\circ$  temporal to the fovea. Since the  $3^\circ$ - $4.5^\circ$  and  $8^\circ$ - $9.5^\circ$  retinal locations are densely packed with somas, the volumes from these locations were cropped to  $0.67^\circ \times 0.67^\circ$  (centered at  $3.75^\circ$ ) and  $0.83^\circ \times 0.83^\circ$  (centered at  $8.5^\circ$ ), respectively, to facilitate manual marking. For brevity, we refer to these three retinal locations as  $3.75^\circ$ ,  $8.5^\circ$ , and  $12.75^\circ$ .

The FDA dataset consisted of  $1.5^\circ \times 1.5^\circ$  volumes at  $12^\circ$  temporal to the fovea,  $2.5^\circ$  superior and inferior of the raphe (for brevity, we refer both locations as  $12^\circ$ ) from five glaucoma patients with hemifield defect (ten volumes; one male and four females,  $56.6 \pm 3.8$  years) and four healthy age-matched subjects (six volumes; three males and one female,  $57.3 \pm 7.3$  years). These volumes were acquired by the multimodal AO retinal imaging system[25] with in retinal tissue resolution of  $2.5 \times 2.5 \times 3.7 \mu\text{m}^3$  (Rayleigh

resolution limit). Volumes from both institutions were the average of 100-250 registered AO-OCT volumes of the same retinal patch. All protocols adhered to the tenets of the Helsinki declaration and were approved by the Institutional Review Boards of Indiana University and the FDA.

#### **4.2.2 Ophthalmic Examination and Glaucoma Diagnosis**

All participants underwent a complete ophthalmological examination. The exam included the measurement of intraocular pressure, a slit lamp examination, dilated fundus examination, determination of axial length with biometry (IOLMaster, Zeiss and Lenstar, Haag Streit), and OCT imaging (Heidelberg Spectralis, Heidelberg, Germany) of the peripapillary retinal nerve fiber layer and macula. The glaucoma subjects were examined by a glaucoma subspecialist and included Humphrey 24-2 and 10-2 visual field (VF) tests (Humphrey Field Analyzer, Carl Zeiss Meditec Inc.). Full clinical records and OCT imaging data were examined to confirm the diagnosis of glaucoma. Automatic OCT retinal layer segmentation of GCL was confirmed and compared to AO imaging locations accounting for the raphe angle. To determine the total deviation (TD) values in the areas corresponding to AO imaged locations, VF 24-2 map was used with accounting for GC displacement[152]. All glaucoma patients included in this work underwent treatment to control intraocular pressure.

#### **4.2.3 Study Design**

We used our datasets to conduct four different experiments to evaluate the performance of our algorithm in: (1) healthy subjects at two trained retinal locations, (2)

healthy subjects at an untrained retinal location (generalizability test), (3) glaucomatous subjects at trained retinal locations, and (4) healthy subjects imaged by two different AO-OCT imagers with training on one and testing on the other (generalizability test).

In the first experiment, we used  $3.75^\circ$  and  $12.75^\circ$  volumes from the IU dataset to train and validate our algorithm and compare against expert-level performance. To attain the gold-standard ground truth GCL soma locations, two expert graders sequentially marked the AO-OCT volumes. After the first expert grader marked the soma locations, the second grader reviewed the labelled somas and corrected these markings as needed. To compare with expert-level performance, we performed an inter-grader variability test in which we obtained a 2<sup>nd</sup> set of manual markings by assigning graders to previously unseen volumes. In total, nine graders were involved in the creation of the manual markings. Details on the assignment of graders to AO-OCT volumes are reported in **Supplementary Table 5**. During training, we used leave-one-subject-out cross-validation to optimally utilize our limited number of labeled datasets. In each fold of cross-validation, we separated one  $12.75^\circ$  volume from the training set as the validation data for monitoring the training process and optimizing the post-processing parameters (see Soma localization and segmentation). In the second experiment, we used the trained CNNs from the first experiment and tested their performances on the  $8.5^\circ$  volumes of the corresponding test subjects without any fine-tuning or modification.

For the third experiment, we used the FDA dataset to evaluate our performance on glaucomatous eyes. To create the gold-standard ground truth, two expert graders

sequentially marked the GCL soma locations. Similar to the IU dataset, the second grader reviewed the first grader’s labels and corrected them as needed. A third independent grader created the “2<sup>nd</sup> Grading” set, which served as the expert-level performance. We trained and optimized our method for both healthy and glaucoma volumes from FDA independently through leave-one-subject-out cross-validation. Next, to test the generalizability of the method between healthy and diseased eyes, we applied the CNN trained on all subjects of one group (healthy or glaucoma) to the other set.

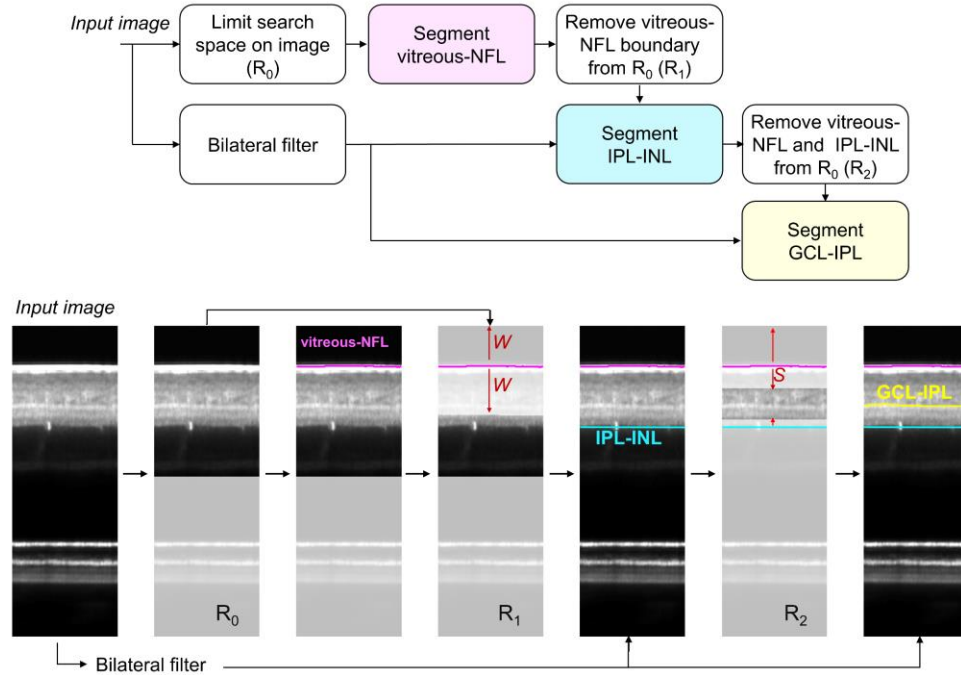
In the last experiment, we tested the generalizability of the method between different imaging systems. We applied the optimized pipeline on data from one device to data from the other device. Specifically, we used the 3.75° and 12.75° volumes from IU and the 12° volumes in the healthy subjects from FDA. Since the devices from these two centers have different voxel sizes (IU:  $0.97 \times 0.97 \times 0.94 \mu\text{m}^3$ , FDA:  $1.5 \times 1.5 \times 0.685 \mu\text{m}^3$ ), we quantified the detection performance with and without test data resized to the training data voxel size. We used cubic interpolation for resizing the AO-OCT volumes.

## **4.2.4 GCL Soma Segmentation**

### **4.2.4.1 Data pre-processing**

We performed retinal layer boundary segmentation as a pre-processing step to narrow the search space for GCL somas. For each volume, we identified the vitreous- NFL and GCL-inner plexiform layer (IPL) boundaries using the graph theory and dynamic programming (GTDP) method described in [112]. Briefly, the GTDP algorithm represents a B-scan image as a graph of connected nodes which are the image pixels. Neighboring

pixels are connected by edges with weights calculated from the image vertical intensity gradients. GTDP then identifies retinal layer boundaries as the minimum weighted path using Dijkstra’s algorithm [153].



**Figure 19: Extraction of the ganglion cell layer (GCL) from AO-OCT volumes.** (Top) Overview of the layer segmentation steps applied to the contrast-enhanced median B-scan image, denoted as the *input image*, and (bottom) illustrative examples for each step of the pipeline. The less transparent areas in  $R_0$ ,  $R_1$ , and  $R_2$  denote regions excluded from the boundary search space.

Using GTDP, we sequentially segmented retinal layer boundaries from the contrast-enhanced median B-scan image (denoted as the *input image* in Figure 19) by limiting the search region using the segmentation result of the previous layer. The schematics in Figure 19 outlines the segmentation steps. First, we set the initial search region ( $R_0$ ) to be the upper half of the image based on the prior knowledge that all layers above the inner nuclear layer (INL) are in this region. Specifically,  $R_0$  is a binary mask with

only the upper half being one. In the case where the outer retina was cropped out from the AO-OCT volume during the manual grading process,  $R_0$  was set to 1 for the entire image. We then segmented the vitreous-NFL. Since the IPL-INL boundary is generally stronger than the GCL-IPL, we identified the IPL-INL prior to segmenting the GCL-IPL. While the vitreous-NFL is easily discernable due to its strong hyper-reflectivity, IPL-INL and GCL-IPL are not as prominent. To accurately segment these two boundaries, we applied a bilateral filter [154] with photometric spread of  $\sigma_r = 1$  and geometric spread of  $\sigma_d = (12, 1.5)$  (horizontal and vertical directions, respectively) to primarily smooth the input image in the horizontal direction while preserving the edges. We then generated a binary mask to narrow the initial search region,  $R_0$ , for a more accurate identification of IPL-INL. We removed the region within  $W \mu\text{m}$  of the vitreous-NFL boundary from  $R_0$  to get the final search region  $R_1$  for IPL-INL. After identifying the IPL-INL with GTDP, we constructed the search region for GCL-IPL ( $R_2$ ). We defined  $R_2$  to be between  $S \mu\text{m}$  below vitreous-NFL and 10 pixels above IPL-INL. We empirically set the parameters as

$$W = \begin{cases} \frac{3}{4}C, & \text{Healthy} \\ \frac{3}{8}C, & \text{Glaucoma} \end{cases}, \quad (1)$$

and,

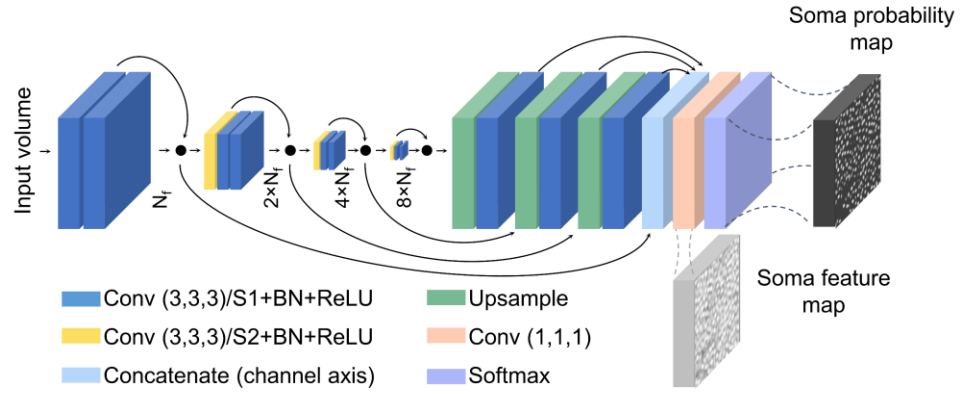
$$S = \begin{cases} \frac{1}{3}C, & \text{Healthy} \\ \frac{1}{10}C, & \text{Glaucoma} \end{cases}, \quad (2)$$

where  $C$  was selected to be  $94 \mu\text{m}$ ,  $70 \mu\text{m}$ , and  $55 \mu\text{m}$  for  $3.75^\circ$ ,  $8.5^\circ$ , and  $12.75^\circ$  recordings, respectively. Based on these boundaries, we extracted volumes extending from the

vitreal-NFL to the GCL-IPL for further analysis. To avoid missing sparsely scattered somas at the GCL-IPL boundary, we kept  $N = 10$  more slices below this boundary.

#### **4.2.4.2 Neural network and training process**

Our neural network is an encoder-decoder CNN with 3D convolutional filters. We designed our CNN such that its encoder path consisted of three down-sampling scales with skip connections to the one-level decoder path (Figure 20). Instead of using max pooling layers to reduce the resolution of the feature maps, we used convolutional layers with stride of 2 for down-sampling and to additionally double the number of feature channels. We incorporated residual learning [155] into the encoder path of our network, which has been shown to facilitate the optimization of deep neural networks. To upscale the feature maps to the input resolution, we used nearest neighbor up-sampling followed by a single convolutional layer. After concatenating the up-sampled feature maps, we used a convolutional layer with two filters to predict features for the background and soma classes. A final softmax layer converted the predicted features into normalized values between 0 and 1 that can be interpreted as class probabilities. All convolutional layers used filters of size  $3 \times 3 \times 3$  voxels (width  $\times$  length  $\times$  depth) and, except the last layer, were followed by batch normalization and rectified linear unit (ReLU) activation.



**Figure 20: Network architecture.** The numbers in parentheses denote the 3D convolutional filter size. The number of filters for each convolutional layer is written on each level.  $N_f = 32$  is the base number of filters. ReLU: rectified linear unit; BN: batch-normalization; Conv: convolution; S: stride.

We created training labels by placing a small sphere (radius of  $2 \mu\text{m}$ ) at each manually annotated GCL soma location. In such labels, most pixels belonged to the background class. We thus used the weighted binary cross-entropy loss to account for this class imbalanced problem. The loss,  $L$ , is defined as

$$L = - \sum_i [w_{pos} y_i \log(p_i) + w_{neg} (1 - y_i) \log(1 - p_i)], \quad (3)$$

where  $y_i$  is the true class label (0 for background, 1 for soma) of voxel  $i$ ,  $p_i$  is the predicted probability for voxel  $i$  to be located on a soma, and  $w_{neg}$  and  $w_{pos}$  are the weights for the background and soma classes, respectively. To reduce the bias towards the background class with higher number of samples, we set  $w_{neg}$  in  $L$  to a lower value than  $w_{pos}$ , thus decreasing its importance in the training process. Specifically, we set  $w_{pos} = 1$  and  $w_{neg} = 0.008$  for the IU dataset and  $w_{pos} = 1$  and  $w_{neg} = 0.002$  for the FDA dataset, which were

determined based on the ratio between the number of voxels in the soma class to the background class in the training labels.

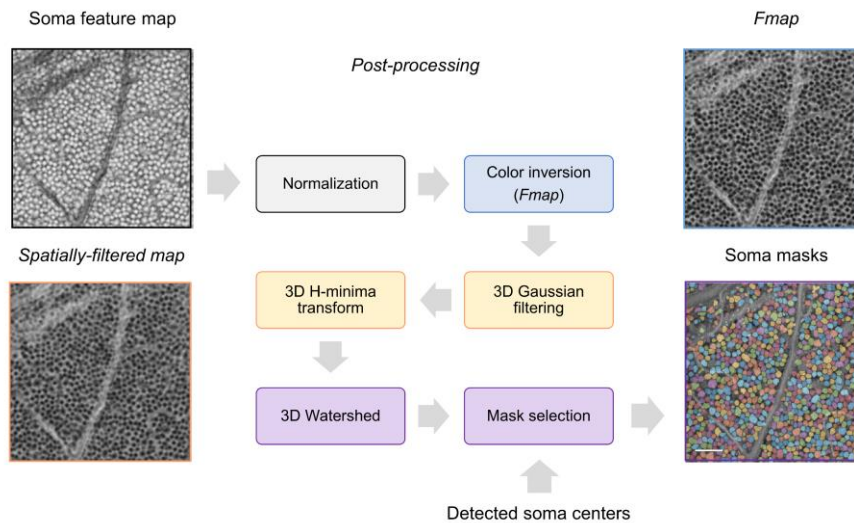
During training, we sampled random batches of two  $120 \times 120 \times 32$  voxel volumes. To improve the generalization abilities of our model, we applied random combinations of rotations ( $90^\circ$ ,  $180^\circ$ , and  $270^\circ$  in the lateral plane) and flips (around all three axis) over the input and label volumes. In addition to these data augmentations, we applied additive zero-mean Gaussian noise with standard deviation of 1.5 to the input volume. We used the Adam optimizer [81] with learning rates of 0.005 and 0.001 for the IU and FDA datasets, respectively. We trained the network for a maximum of 120 epochs with 100 training steps per epoch, during which the loss function converged in all our experiments. We used the network weights that resulted in the highest detection score (explained in *4.2.5 Performance Evaluation*) on the validation data for further analysis.

During the CNN training on the  $3.75^\circ$  and  $12.75^\circ$  volumes, we accounted for the inhomogeneous presence of the larger parasol GCs between these two locations by exposing the CNN to the  $12.75^\circ$  volumes more than the  $3.75^\circ$  location. Specifically, we set the probability of selecting the  $12.75^\circ$  volumes to be five times higher than the  $3.75^\circ$  volumes.

#### **4.2.4.3 Soma localization and segmentation**

As the final step of our method, we post-processed the network's predictions and final feature map (features before the Softmax layer) to localize GCL somas and obtain the segmentation masks, respectively. We input AO-OCT volumes into the trained network

using a  $256 \times 256 \times 32$  voxel sliding window with step size equal to half the window size. In the overlapping regions, we averaged the predictions (and features). Additionally, we used test-time-augmentation (TTA) to enhance soma detection and segmentation. This step consisted of averaging network predictions (and features) for eight rotations and reflections in the *en face* plane of the input volume. Next, we applied median filter of size  $3 \times 3 \times 3$  to the prediction maps to remove spurious maxima. We then located somas from the filtered maps by finding points that were local maxima in a  $3 \times 3 \times 3$  ( $3 \times 3 \times 7$  for FDA) voxel window and had probability values greater than  $T$ . We used the validation data to find the value of  $T$  that maximized the detection performance (see 4.2.5 Performance Evaluation).



**Figure 21: Unsupervised segmentation of GCL somas using the CNN’s learned features. The colored boxes correspond to steps with matching colors. Scale bar: 50 μm.**

To segment somas, we used the network’s final feature volumes belonging to the soma class (Figure 20 and Figure 21). After normalizing the features to the range between

0 and 1 and inverting the intensities (0s became 1s and vice versa), denoting the result as  $Fmap$ , we used the 3D watershed algorithm to obtain the masks for individual somas. To prevent over-segmentation by the watershed algorithm, we first smoothed  $Fmap$  with a 3D Gaussian filter and applied the H-minima transform using MATLAB's (MathWorks)  $imhmin$  function with parameter 0.01. We set the filter's standard deviation to (1.5, 1) pixels (*en face* and axial planes, respectively) and (0.4, 0.7) pixels for the IU and FDA datasets, respectively. Finally, we removed voxels with intensity values greater than 0.6 in the filtered  $Fmap$  volume from the set of watershed masks. We estimated soma diameter as the diameter of an equivalent circle with an area equal to that of the soma's *en face* mask image. In practice, we used information from one C-scan below to one C-scan above the predicted soma center to obtain more accurate estimates. Eye length was used to scale the retinal images from degrees to millimeters [156].

#### 4.2.5 Performance Evaluation

After training, we applied the network to the hold-out data for testing the performance. We evaluated the detection performance of our method by comparing the results with the gold-standard labels. We assessed the detection performance by using recall, precision, and  $F_1$  scores.

To determine the true positive somas, we used the Euclidean distance between the automatically found and the manually marked somas. Each manually marked soma was matched to its nearest automatic soma if the distance between them was smaller than  $D$ . We set the value for  $D$  to half of the previously reported mean GCL soma diameters in

healthy eyes for each retinal location. For the glaucoma case, we used 0.75 times the median spacing between manually marked somas. This yielded  $D$  values of 5.85  $\mu\text{m}$  and 8.78  $\mu\text{m}$  for the 3.75° and 12°-12.75° volumes for healthy subjects, respectively, and 10.78  $\mu\text{m}$  for the 12° volumes from glaucoma patients. To remove border artifacts, we did not analyze somas within 10 pixels from the volume edges. For inter-observer variability, we compared the markings of the 2<sup>nd</sup> grading to the gold-standard markings in the same way.

In addition to detection performance scores, we compared our estimated cell densities to the gold-standard values. We measured cell density by dividing the cell count to the image area after accounting for large blood vessels and image edges. We did not differentiate GCs from displaced amacrine cells, which represent 22% of the somas in the GCL at 13° in healthy eyes [157]. We did not offset our counts with this value to facilitate comparison between healthy and glaucomatous eyes, for which we cannot determine the level of amacrine cell degeneration. Finally, for evaluating the segmentation accuracy, we compared our predicted soma diameters to data from previous *in vivo* semi-automatic measurements [24, 111] and histological studies [157-161].

## ***4.3 Results***

### **4.3.1 Achieving Expert Performance on Healthy Subjects and Generalizing to an Unseen Retinal Location**

We first quantified the performance of our method using AO-OCT volumes taken from healthy individuals using Indiana University's imaging system. In the human retina, GCL somas reach a maximum density and cell stack depth near 3°-4.5°, where these GCs

project to the densely packed cone photoreceptors at the foveal center [24, 157, 162]. At increasing eccentricities from this peak, the cell density and GCL thickness monotonically decrease, with GCs eventually arranged in a monolayer around 12°-13° [24, 157, 162]. The GC soma size also varies with retinal eccentricity, with some GC types varying more than others. The two primary subtypes of GCs are the midget and parasol cells; parasol GC somas are generally larger than midget GC somas (e.g. at 12°-13°) but are increasingly similar and smaller in size closer to the fovea. At 3°, the two GC types are effectively identical in size and thus indistinguishable based on size [159, 160, 163].

Using the characteristically different 3.75° and 12.75° volumes (in terms of GCL soma sizes and size distributions), we trained our CNN through leave-one-subject-out cross-validation. The GCL soma detection performance of our automated method compared to the gold-standard manual markings and an inter-observer test are summarized in Table 7. Our method surpassed or was on par with expert performance in detecting GCL somas ( $p$ -values = 0.008 and 0.078 for 3.75° and 12.75° volumes, respectively; two-sided Wilcoxon signed rank test over  $F_1$  scores of  $n = 8$  subjects).

Next, with the aim of evaluating the generalizability of the method to a previously unseen location, we applied the optimized method to a new retinal region. To this end, we used the trained networks and optimized parameters on the 3.75° and 12.75° volumes, and we tested the performance on the 8.5° data. On the unseen 8.5° volumes, our method performed at a level indistinguishable from those of the former two retinal locations in terms of the  $F_1$  score (Table 7) and was on par with expert performance ( $p$ -value = 0.063;

two-sided Wilcoxon signed rank test over  $n = 5$  subjects). To provide a more complete picture of the performance, the average precision-recall curves of our method compared to the average expert grading performance are shown in Figure 22a. The average curves were obtained by taking the mean of the precision and recall values of all the trained networks at the same threshold value. When matched to have the same average precision as the expert grader, the average recall scores of our method were 0.16, 0.08, and 0.08 higher at the  $3.75^\circ$ ,  $8.5^\circ$ , and  $12.75^\circ$  locations, respectively. Our method’s generalizability and expert-level detection performance persisted even if we used other intensity transformations on the GCL volumes prior to network training or disregarded test-time-augmentation (Figure 22a and Table 8).

**Table 7: GCL soma detection performance scores for the IU dataset. Scores are reported as mean (standard deviation) calculated across  $n = 8$  subjects for  $3.75^\circ$  and  $12.75^\circ$  and a subset of  $n = 5$  subjects for the  $8.5^\circ$  location. The  $8.5^\circ$  volumes were not involved in the training or optimization process. At each retinal eccentricity, the higher detection performance score (CNN vs 2<sup>nd</sup> Grading) is written in bold**

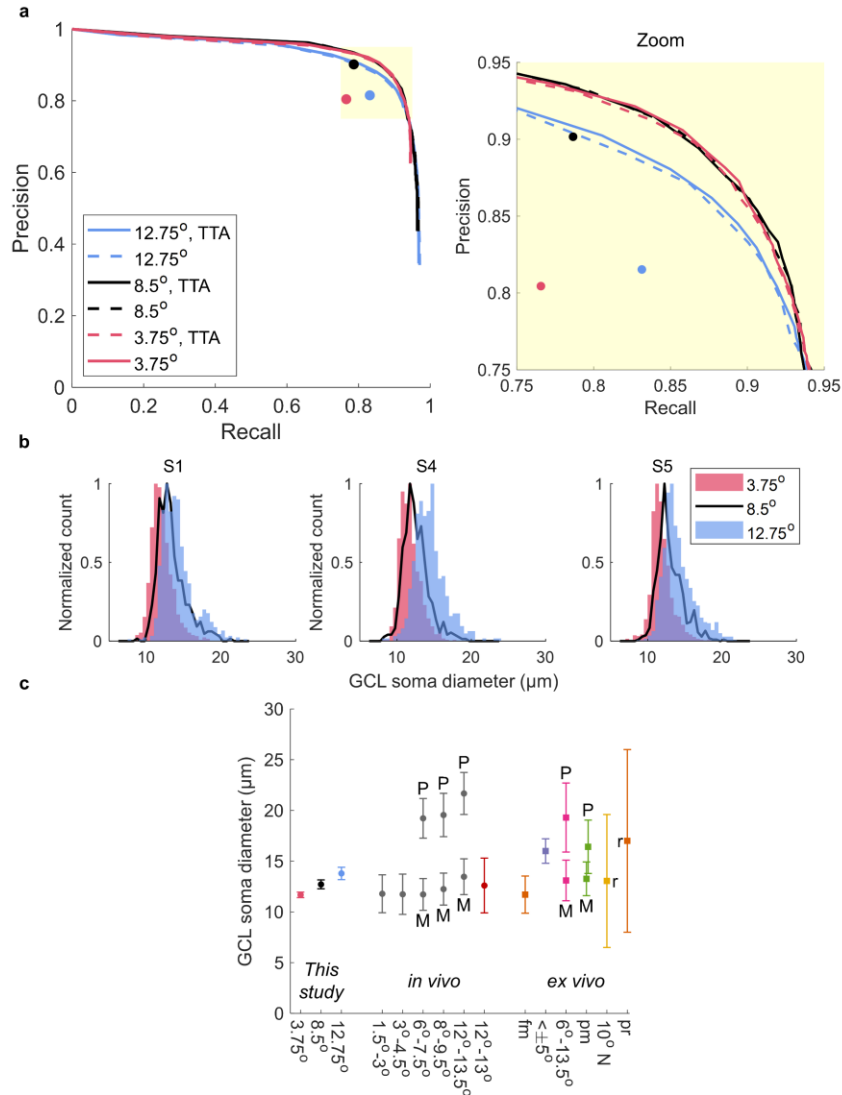
Eccentricity	Method	GCL Soma Detection		
		Recall	Precision	F <sub>1</sub>
$3.75^\circ$	CNN	<b>0.88 (0.09)</b>	<b>0.87 (0.06)</b>	<b>0.87 (0.04)</b>
	2 <sup>nd</sup> Grading	0.77 (0.16)	0.80 (0.08)	0.77 (0.06)
$12.75^\circ$	CNN	<b>0.88 (0.07)</b>	<b>0.85 (0.06)</b>	<b>0.87 (0.04)</b>
	2 <sup>nd</sup> Grading	0.83 (0.10)	0.82 (0.12)	0.81 (0.05)
$8.5^\circ$	CNN	<b>0.90 (0.05)</b>	0.85 (0.04)	<b>0.87 (0.02)</b>
	2 <sup>nd</sup> Grading	0.79 (0.02)	<b>0.90 (0.03)</b>	0.84 (0.02)

**Table 8: Effect of intensity normalization and test-time-augmentation on detection performance for IU’s dataset. Scores are reported as mean  $\pm$  standard deviation for  $F_1$  (recall, precision), calculated across  $n = 8$  subjects for  $3.75^\circ$  and  $12.75^\circ$  and  $n = 5$  subjects for the  $8.5^\circ$  locations.**

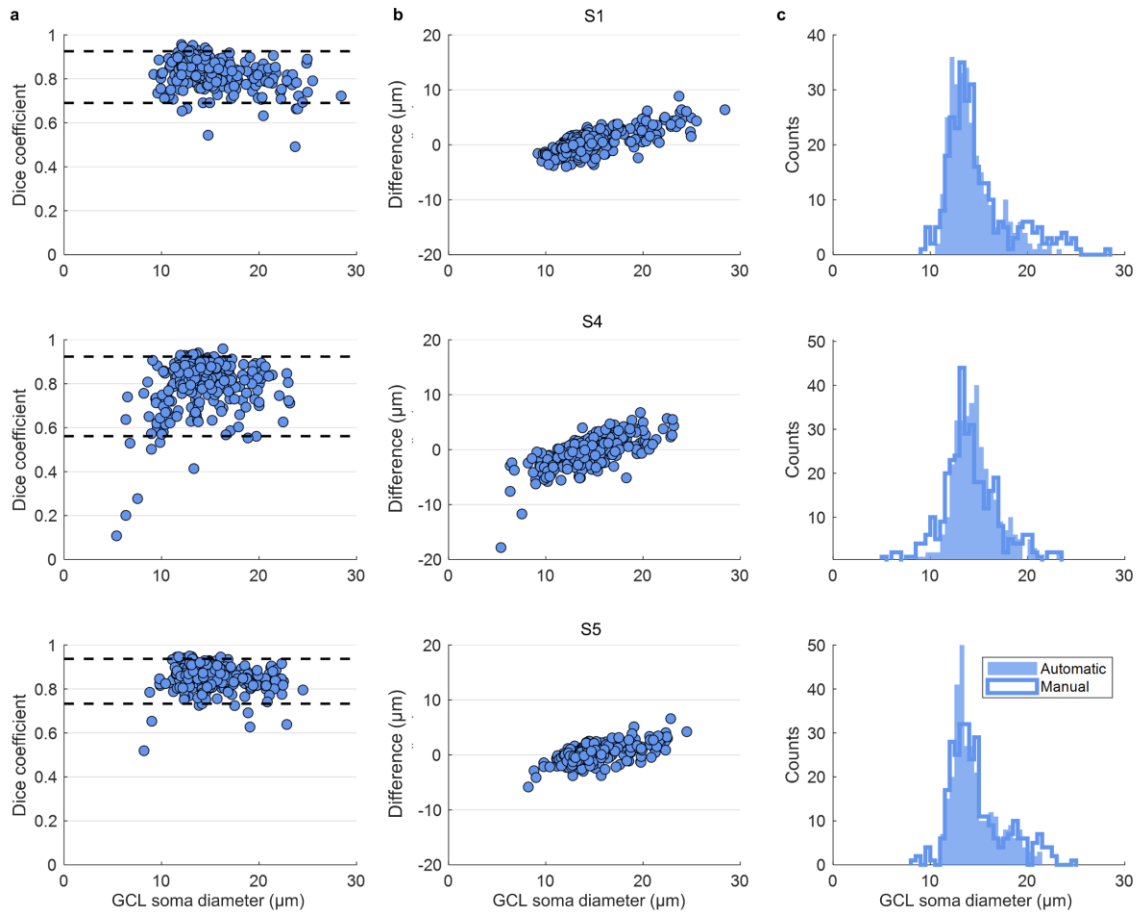
Intensity normalization	Test-Time-Augmentation		
	Eccentricity	Yes	No
Identity	3.75°	0.87 $\pm$ 0.04 (0.88 $\pm$ 0.09, 0.87 $\pm$ 0.06)	0.87 $\pm$ 0.04 (0.87 $\pm$ 0.10, 0.88 $\pm$ 0.05)
	8.5°	0.87 $\pm$ 0.02 (0.90 $\pm$ 0.05, 0.85 $\pm$ 0.04)	0.88 $\pm$ 0.01 (0.89 $\pm$ 0.04, 0.87 $\pm$ 0.02)
	12.75°	0.87 $\pm$ 0.04 (0.88 $\pm$ 0.07, 0.85 $\pm$ 0.06)	0.86 $\pm$ 0.05 (0.88 $\pm$ 0.06, 0.85 $\pm$ 0.07)
Whiten	3.75°	0.85 $\pm$ 0.06 (0.85 $\pm$ 0.14, 0.87 $\pm$ 0.07)	0.86 $\pm$ 0.05 (0.86 $\pm$ 0.13, 0.88 $\pm$ 0.05)
	8.5°	0.88 $\pm$ 0.02 (0.89 $\pm$ 0.03, 0.86 $\pm$ 0.02)	0.88 $\pm$ 0.01 (0.90 $\pm$ 0.04, 0.86 $\pm$ 0.02)
	12.75°	0.85 $\pm$ 0.06 (0.86 $\pm$ 0.08, 0.85 $\pm$ 0.08)	0.88 $\pm$ 0.06 (0.87 $\pm$ 0.06, 0.85 $\pm$ 0.08)

Using the method’s soma segmentation masks, we estimated the GCL soma diameters in the *en face* plane. The histograms of soma diameter values in Figure 22b reflect the trend of gradual increase in soma size from  $3.75^\circ$  to  $12.75^\circ$ , which is consistent with the retinal GC populations at these locations. The comparison between the estimated diameters with reported values from literature (Figure 22c) indicates that our predicted values are in line with values from *ex vivo* histological [157-161] and *in vivo* semi-automatic [24, 111] measurement studies. To further validate the segmentation accuracy, we manually segmented 300-340 randomly selected GCL somas from the  $12.75^\circ$  volumes of three subjects. For each soma, we used the *en face* plane at which the soma center was located. The automatic segmentation masks agreed with the manual masks (mean (95%

confidence interval) of Dice similarity coefficients = 0.83 (0.82, 0.84), 0.81 (0.79, 0.82), and 0.86 (0.85, 0.87) for subjects S1, S4, and S5, respectively; Figure 23a).



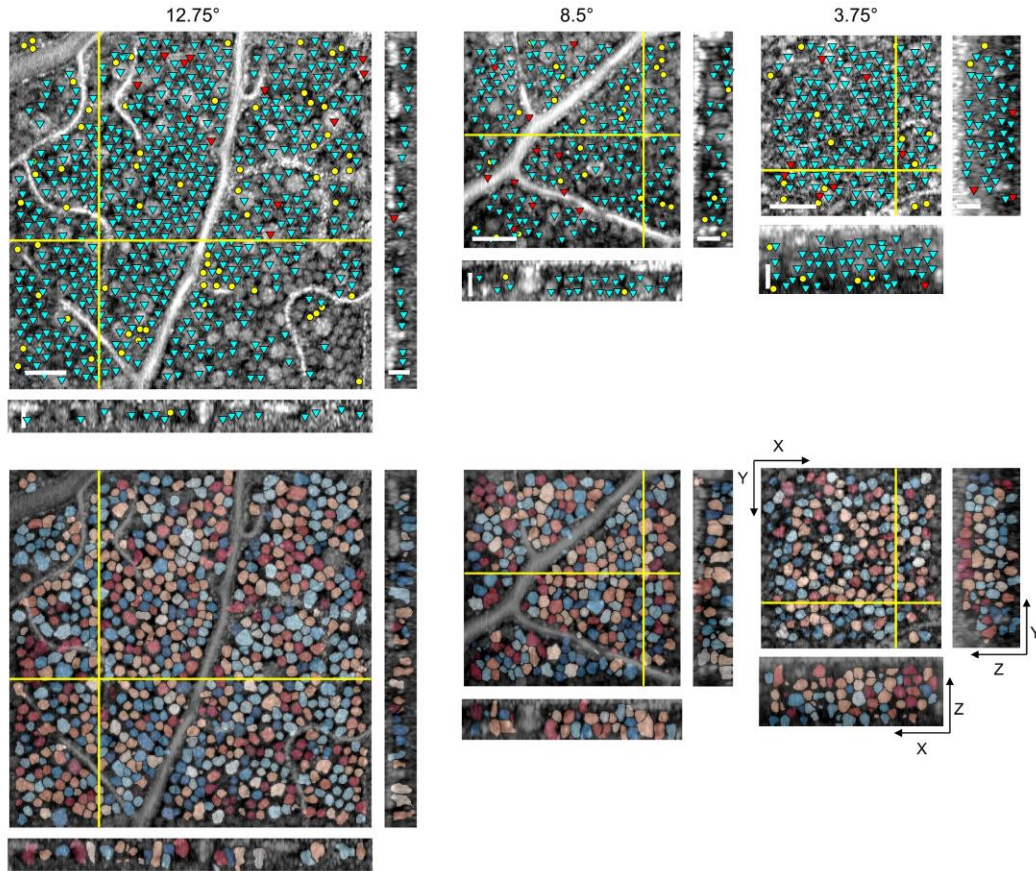
**Figure 22: Our method achieved expert-level performance across different retinal locations on the IU dataset. a, Average precision-recall curves of our method compared to average expert grader performances (circle markers). Each plotted curve is the average of  $n = 8$  and 5 curves at the same threshold values for the 3.75°/12.75° and 8.5° data, respectively. TTA: test-time-augmentation. b, Soma size distributions for three subjects across different retinal locations. c, GCL soma diameters across all subjects compared to previously reported values. Error bars denote one standard deviation unless labeled with “r” to denote range of values. P: parasol GCs, M: midget GCs, fm: foveal margin, pm: papillomacular, pr: peripheral retina.**



**Figure 23: Comparison between automatic and manual GCL soma segmentations. a, Dice similarity coefficients between the automatic and manual masks, b, error in soma diameter measurements, and c, histogram of measured diameters for subjects S1, S4, and S5. Markers denote individual somas ( $n = 300-343$  somas). Error is defined as the automatically determined soma diameter minus the automatically measured diameter. Dice coefficients and soma diameters were calculated based on the binary masks at the *en face* plane where each soma's center was located. Dashed black lines in a are the 95% data intervals.**

Example results with their comparison to the gold-standard manual markings are illustrated in Figure 24. The cyan, red, and yellow markers indicate correctly identified (true positive), missed (false negative), and incorrectly identified (false positive) somas, respectively. The average prediction times were  $2.0 \pm 0.5$ ,  $1.3 \pm 0.1$ , and  $3.2 \pm 0.5$  minutes/volume for the  $3.75^\circ$ ,  $8.5^\circ$ , and  $12.75^\circ$  data, respectively. These prediction speeds

were at least two orders of magnitude faster than that of manual grading, which took 7-8 hours/volume.



**Figure 24: Illustrative results on the IU dataset.** Soma detection and segmentation results on volumes from one subject. *En face* (XY) and cross-sectional (XZ and YZ) slices illustrate (top) soma detection results compared to the gold-standard manual markings and (bottom) overlay of soma segmentation masks, with each soma represented by a randomly assigned color. Cyan, red, and yellow markers denote true positives, false negatives, and false positives, respectively. Only somas with centers located within  $5\ \mu\text{m}$  from the depicted slices in the top row are marked. The intensities of AO-OCT images are shown in log-scale. Scale bars:  $50\ \mu\text{m}$  and  $25\ \mu\text{m}$  for the *en face* and cross-sectional slices, respectively.

### 4.3.2 Achieving Expert Performance on Glaucoma Patients

To demonstrate the ability of our method to process diseased eyes, we next applied it to AO-OCT volumes taken from glaucomatous eyes. Volumes from healthy and

glaucoma subjects at 12° temporal to the fovea were acquired with FDA’s imaging system. We whitened each extracted NFL+GCL volume by subtracting its mean and dividing by its standard deviation. We then trained our method separately on the two groups of subjects and validated the performances through leave-one-subject-out cross-validation.

**Table 9: GCL soma detection performance scores on the FDA dataset. Scores are reported as mean (standard deviation) calculated across  $n = 6$  healthy and  $n = 10$  glaucoma AO-OCT volumes. For each group, the higher detection performance score (CNN vs 2<sup>nd</sup> Grading) is written in bold.**

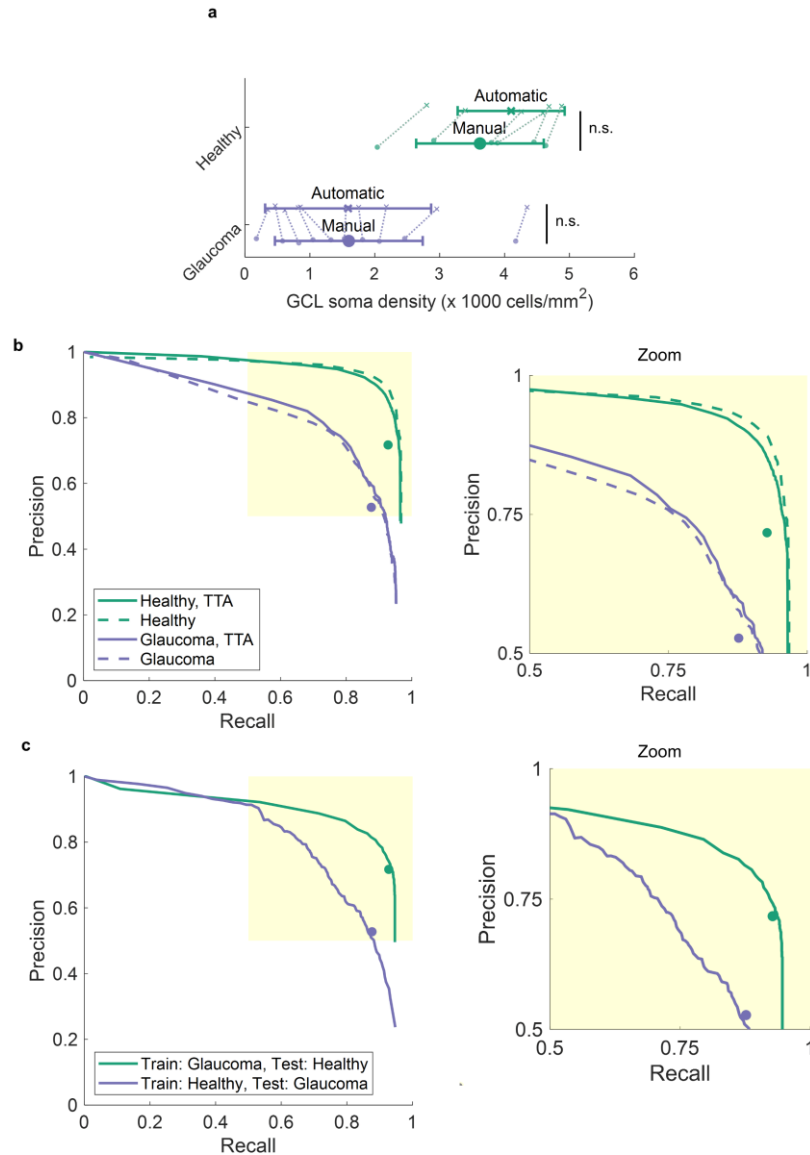
Group	Method	GCL Soma Detection		
		Recall	Precision	F <sub>1</sub>
Healthy	CNN	0.90 (0.04)	<b>0.78 (0.07)</b>	<b>0.84 (0.05)</b>
	2 <sup>nd</sup> Grading	<b>0.93 (0.02)</b>	0.72 (0.09)	0.81 (0.06)
Glaucoma	CNN	0.75 (0.14)	<b>0.78 (0.15)</b>	<b>0.75 (0.11)</b>
	2 <sup>nd</sup> Grading	<b>0.88 (0.07)</b>	0.53 (0.16)	0.64 (0.13)

The method’s automatically estimated cell densities were similar to the gold-standard values for both groups ( $p$ -values = 0.125 and 1 across  $n = 4$  and 5 healthy and glaucoma subjects, respectively; two-sided Wilcoxon signed rank test; Figure 25a). Table 9 summarizes the method’s detection results compared to the gold-standard manual markings and the inter-observer test results. For both groups, our results were on par with expert performance based on the average F<sub>1</sub> scores of each subject ( $p$ -values = 0.125 and 0.063 over  $n = 4$  and 5 healthy and glaucoma subjects, respectively; two-sided Wilcoxon signed rank test).

Figure 25b depicts the average precision-recall curves of our trained networks compared to the average expert grader performance; at the same level of average expert grader precision, our method achieved 0.04 and 0.03 higher average recall scores for the healthy and glaucoma subjects, respectively. Our method achieved high detection scores even without data whitening or test-time-augmentation (Figure 25b and Table 10). Moreover, the method retained expert-level performance when tested on a group not used during training (Figure 25c and Table 11), reflecting its generalizability between healthy and diseased eyes. Figure 26 illustrates comparisons between our detection results with the gold-standard manual markings and the volumetric soma segmentation results for a healthy and a glaucoma subject.

**Table 10: Effect of intensity normalization and test-time-augmentation on detection performance for FDA’s dataset. Training and optimizations were done independently for the two groups of subjects.**

		Test-Time-Augmentation	
		Yes	No
Intensity normalization	Dataset		
	Healthy	$0.76 \pm 0.17$ ( $0.76 \pm 0.24, 0.83 \pm 0.05$ )	$0.77 \pm 0.12$ ( $0.78 \pm 0.21, 0.80 \pm 0.06$ )
Identity	Glaucoma	$0.64 \pm 0.19$ ( $0.59 \pm 0.25, 0.74 \pm 0.16$ )	$0.61 \pm 0.20$ ( $0.57 \pm 0.27, 0.71 \pm 0.20$ )
	Healthy	$0.84 \pm 0.05$ ( $0.90 \pm 0.04, 0.78 \pm 0.07$ )	$0.83 \pm 0.05$ ( $0.87 \pm 0.06, 0.80 \pm 0.05$ )
Whiten	Glaucoma	$0.75 \pm 0.11$ ( $0.75 \pm 0.14, 0.78 \pm 0.15$ )	$0.73 \pm 0.12$ ( $0.78 \pm 0.14, 0.72 \pm 0.16$ )



**Figure 25: Results on FDA's healthy and glaucoma subjects.** **a**, Automatically estimated cell densities (cross markers) agreed with the gold-standard manual values (circle markers) for both groups. Corresponding points from the automatic and manual sets are connected. Error bars denote one standard deviation. P-values = 0.125 and 1 (two-sided Wilcoxon signed rank test with  $n = 4$  and  $5$  subjects) for the healthy and glaucoma groups, respectively. n.s: not significant. **b**, Average precision-recall curves of our method compared to average expert grader performances (circle markers). **c**, Average precision-recall curves of our method trained on one group of subjects and tested on the other compared to the average expert grader. Each plotted curve is the average of  $n = 6$  and  $10$  curves for the healthy and glaucoma volumes, respectively.

**Table 11: Generalizability test between groups of subjects from the FDA dataset. Networks were trained with whitened volumes, and predictions were made with test-time-augmentation. Validation data denote volumes used to determine the post-processing parameters. For tests with inter-group test and validation data, parameters were optimized on a randomly selected dataset from the validation group. Leave-one-subject-out cross-validation was used for tests with intra-group test and validation data.**

		Validation Data	
Test Data	Training Data	Healthy	Glaucoma
Healthy	Healthy	$0.84 \pm 0.05$ ( $0.90 \pm 0.04, 0.78 \pm 0.07$ )	-
	Glaucoma	$0.84 \pm 0.03$ ( $0.88 \pm 0.05, 0.81 \pm 0.05$ )	$0.84 \pm 0.03$ ( $0.89 \pm 0.04, 0.81 \pm 0.05$ )
Glaucoma	Healthy	$0.58 \pm 0.22$ ( $0.47 \pm 0.27, 0.92 \pm 0.07$ )	$0.71 \pm 0.11$ ( $0.71 \pm 0.18, 0.75 \pm 0.11$ )
	Glaucoma	-	$0.75 \pm 0.11$ ( $0.75 \pm 0.14, 0.78 \pm 0.15$ )

Using the soma segmentation masks, we estimated cell diameters in the *en face* plane. As illustrated in Figure 27a, the estimated soma diameters on the healthy cohort agrees with the estimates from the IU dataset and previous studies at 12°-13°. The results also reflect an increase of 2.88  $\mu\text{m}$  ( $p$ -value = 0.03, Wilcoxon rank sum test with  $n = 5$  glaucoma and 4 healthy subjects, respectively) in the average soma size of glaucoma subjects compared to healthy individuals, which is in line with recently reported semi-automatic measurements [164].

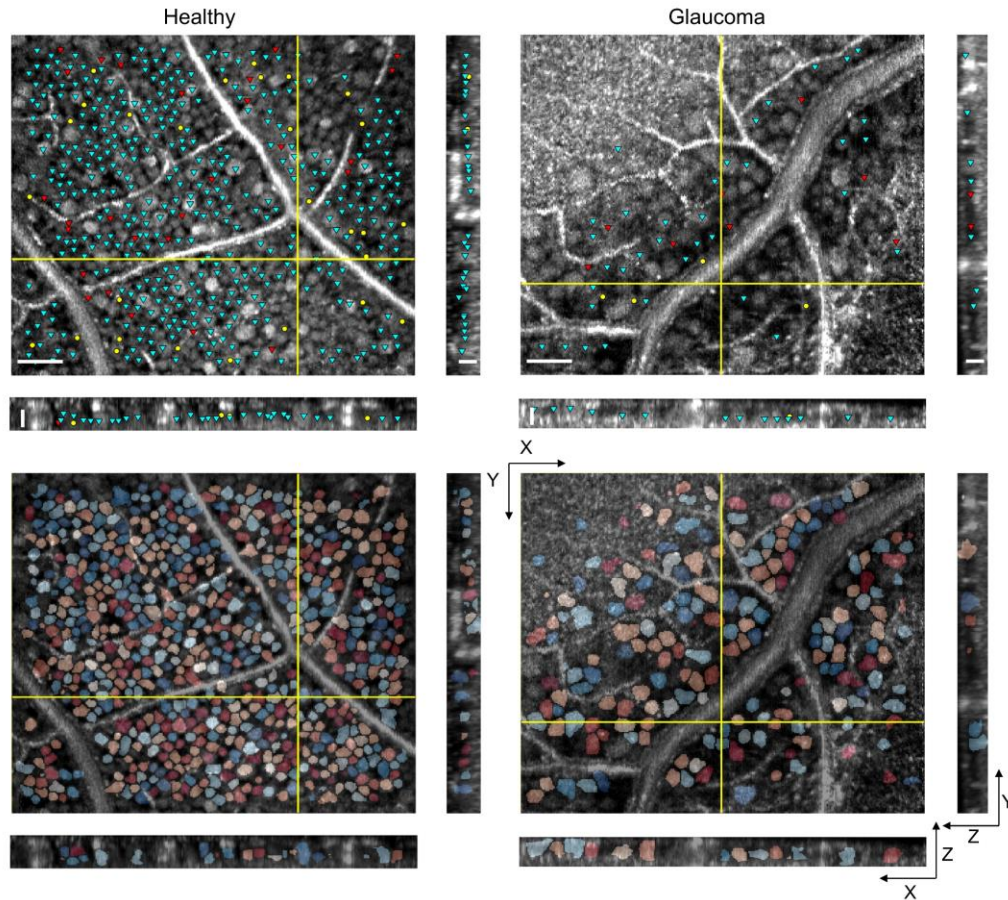
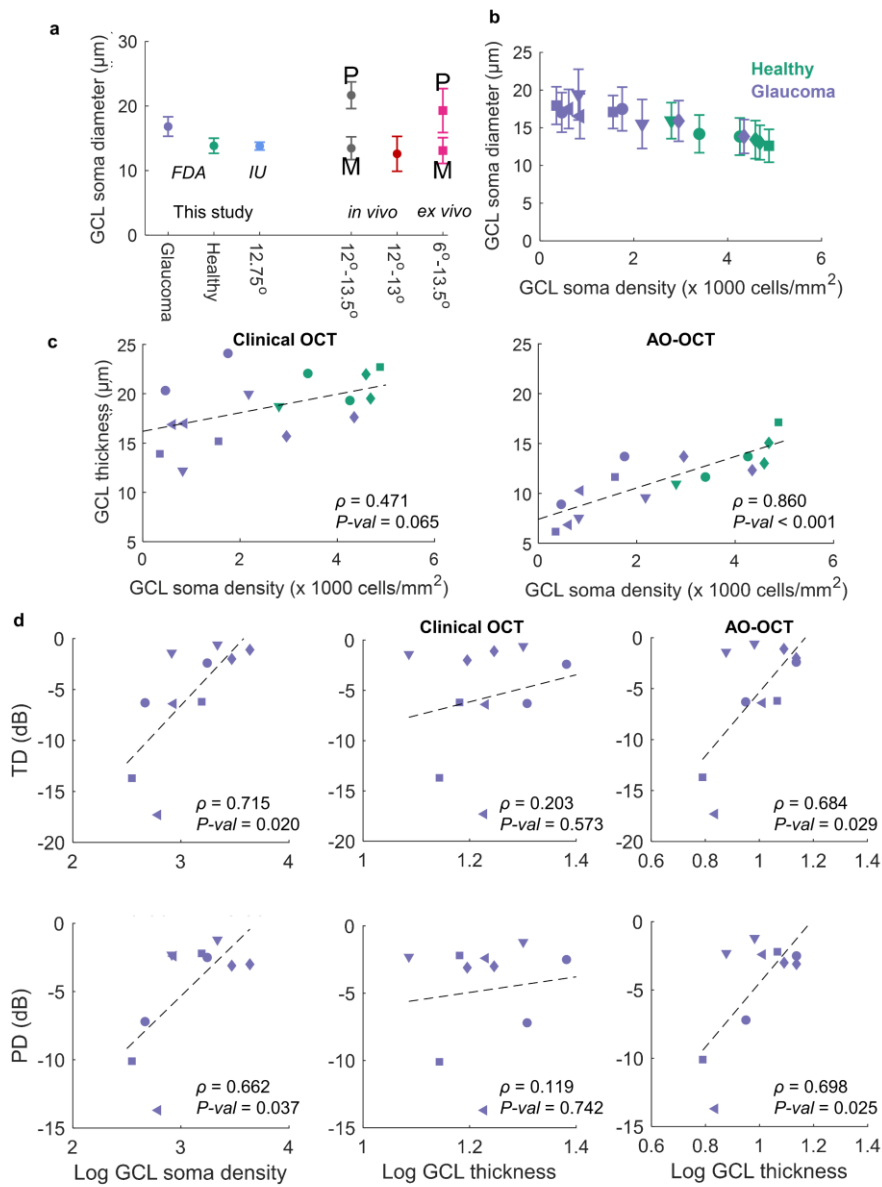


Figure 26: Illustrative results on FDA's dataset. *En face* (XY) and cross-sectional (XZ and YZ) slices illustrate (top) soma detection results compared to the gold-standard markings and (bottom) overlay of individual segmentation masks, with each soma represented by a randomly assigned color. Cyan, red, and yellow markers denote true positives, false negatives, and false positives, respectively. Only somas with centers located within  $5\ \mu\text{m}$  from the depicted slices in the top row are marked. Scale bars:  $50\ \mu\text{m}$  and  $25\ \mu\text{m}$  for the *en face* and cross-sectional slices, respectively.

### 4.3.3 Structural and Functional Characteristics of Glaucomatous Eyes Differ from Control Eyes

Using our automatic quantification method, we next examined the differences in cellular-level characteristics and clinical data of glaucomatous eyes compared to healthy eyes. Figure 27b illustrates the GCL soma size against cell densities, reflecting that glaucoma subjects exhibited lower number of cells that were larger than somas in the

healthy group. As shown in Figure 27c, the automatically determined cell densities strongly correlated with GCL thicknesses measured from AO-OCT images (Pearson correlation coefficient,  $\rho = 0.860$ ,  $p$ -value  $< 0.001$ ), whereas they did not correlate with thickness values obtained from clinical OCT images ( $\rho = 0.471$ ,  $p$ -value = 0.065). For analysis of functional measures in glaucoma subjects, we used the total deviation (TD) and pattern deviation (PD) values (in decibels) provided by the Humphrey 24-2 visual field test. TD and PD values represent the local loss in sensitivity and focal depressed areas compared with age-matched controls, respectively. PD values account for any depression of the hill of vision that might be caused by cataracts, vitreous hemorrhages, or other diffuse media opacities. When comparing these local functional measures with the local structural characteristics (Figure 27d), the soma density in log-scale correlated with TD ( $\rho = 0.715$ ,  $p$ -value = 0.02) and PD ( $\rho = 0.662$ ,  $p$ -value = 0.037). The log-scale GCL thickness from AO-OCT images also correlated with these measures ( $\rho = 0.684$  and  $0.698$ ,  $p$ -values = 0.029 and 0.025 for TD and PD, respectively), while the measurements from clinical OCT lacked correlation with the functional data ( $\rho = 0.203$  and  $0.119$ ,  $p$ -values = 0.573 and 0.742 for TD and PD, respectively). GCL thickness and soma density, when used together as independent variables, increased the correlation with both TD and PD (coefficient of multiple correlation,  $R = 0.735$  and  $0.715$ , respectively). Adding the GCL soma diameters as an additional independent variable further increased the correlation with the functional measures ( $R = 0.760$  and  $0.756$  for TD and PD, respectively).



**Figure 27: Structural and functional characteristics of glaucomatous eyes compared to controls. a, GCL soma diameters across all subjects compared to values reported in the literature. b, Automatic cell densities and average diameters for all volumes. c, GCL thickness versus soma densities for all volumes measured with clinical OCT and AO-OCT images. d, Total deviation (TD) and pattern deviation (PD) measurements versus cell densities and GCL thickness values in log-scale for glaucoma subjects.  $\rho$ : Pearson correlation coefficient. In b-d, each subject is shown with a different marker shape. Some subjects had imaging at both  $2.5^\circ$  superior and inferior of the raphe at  $12^\circ$ . Error bars in a-b denote one standard deviation.**

### 4.3.4 Generalizing Between Imaging Devices

The aforementioned results were obtained from two imaging systems with different scan and sampling characteristics by training models separately for each device. The voxel size of AO-OCT volumes imaged with IU’s device were  $0.97 \times 0.97 \times 0.94 \mu\text{m}^3$  (width  $\times$  length  $\times$  depth), compared to  $1.5 \times 1.5 \times 0.685 \mu\text{m}^3$  for volumes acquired by FDA. To evaluate the generalizability of the method between these devices, we applied the trained and optimized method on data from one device to volumes acquired by the other system (Table 12).

**Table 12: Generalizability of trained method across different AO-OCT imaging systems with different scan characteristics. Detection scores are reported as mean  $\pm$  standard deviation for  $F_1$  (recall, precision) across  $n = 16$  and  $6$  volumes for IU and FDA, respectively. The IU dataset consisted of the  $3.75^\circ$  and  $12.75^\circ$  locations. Networks were trained with whitened volumes, and predictions were made with test-time-augmentation. Resizing denotes interpolation of the test volumes to the voxel size of the training dataset before passing them to the networks.**

Test Data	Training Data	Resizing Test Volumes	
		Yes	No
IU	IU	-	$0.85 \pm 0.06$ ( $0.86 \pm 0.11, 0.86 \pm 0.07$ )
	FDA Healthy	$0.86 \pm 0.05$ ( $0.83 \pm 0.08, 0.90 \pm 0.05$ )	$0.64 \pm 0.16$ ( $0.52 \pm 0.17, 0.92 \pm 0.04$ )
FDA Healthy	IU	$0.75 \pm 0.10$ ( $0.93 \pm 0.07, 0.63 \pm 0.12$ )	$0.55 \pm 0.16$ ( $0.43 \pm 0.17, 0.83 \pm 0.06$ )
	FDA Healthy	-	$0.84 \pm 0.05$ ( $0.90 \pm 0.04, 0.78 \pm 0.07$ )

To this aim, we used IU’s  $3.75^\circ$  and  $12.75^\circ$  volumes and FDA’s healthy data at  $12^\circ$ . After resizing the test volumes to the same voxel size as the training data, the detection

performance of the inter-device testing scheme (mean  $\pm$  SD  $F_1$  scores of FDA train, IU test:  $0.86 \pm 0.05$ , and IU train, FDA test:  $0.75 \pm 0.10$ ) was similar to that of the intra-device framework (IU:  $0.85 \pm 0.06$ , and FDA:  $0.84 \pm 0.05$  with  $p$ -values = 0.844 and 0.125 over  $n = 8$  and 4 subjects, respectively; two-sided Wilcoxon sign rank test on the average  $F_1$  scores of each subject) without additional parameter optimization. Without test volume resizing and parameter tuning, the trained method on one device could not necessarily generalize to the other imager ( $F_1$  scores of FDA train, IU test:  $0.64 \pm 0.16$ , and IU train, FDA test:  $0.55 \pm 0.16$  with  $p$ -values = 0.008 and 0.125 over  $n = 8$  and 4 subjects, respectively; two-sided Wilcoxon sign rank test).

#### ***4.4 Discussion***

Our work provides a first step toward automatic quantification of GCL somas from AO-OCT volumes. We developed a weakly-supervised deep learning-based method to automatically segment individual somas without manual segmentation masks, which are expensive to acquire. We trained the CNN using expert markings of soma locations. Using the trained localization network, we devised an unsupervised method to segment individual GCL somas. We trained and validated our method on healthy and glaucoma subjects and two separate imaging devices across multiple retinal locations. Compared to manual marking of volumes, which took between 7-8 hours/volume, our method was at least two orders of magnitude faster with a speed of less than 3 minutes/volume.

Our method achieved high detection performance regardless of retinal eccentricity, imaging device, or presence of pathology. The ability of our technique to

identify GCL somas matched or exceeded that of experts and generalized to an unseen retinal location, between healthy and glaucoma subjects, and between devices. Additionally, our method's soma segmentation masks agreed with manual segmentation masks and the mean and range of estimated soma diameter values were comparable to the reported values from previous *ex vivo* and *in vivo* measurement studies.

Our estimated soma diameters differed from previous studies in two aspects. First, the inter-subject standard deviation (SD) of mean soma diameters for individuals involved in this study (measured automatically; error bars in Figure 22c) were smaller than the values reported by Liu *et al.* [24]. This dissimilarity could be due to differences between the soma diameter measurement approaches taken by us and this study. In our work, we approximated soma diameters with the diameter of a circle with the same area as the soma segmentation mask (for both manual and automatic). Whereas Liu *et al.* [24] defined soma diameter as twice the distance between the manually marked soma center and the minimum in the circumferential intensity averaged trace around the soma center. The other *in vivo* diameter measurement study by Rossi *et al.* [111] has reported the mean and SD over the set of all manually segmented somas of two human subjects, which cannot be directly compared to our inter-subject SD of mean soma diameters reported in Figure 22c. In addition, Rossi *et al.* measured GCL soma diameters from AO-SLO images, which are different from AO-OCT images in terms of image quality. The inherent inter- and intra- variability of human graders in marking images due to the subjective nature of the task, as has been demonstrated for OCT and AO-SLO images [132, 165, 166], could

also contribute to the higher SD values of previous studies. In contrast, our automatic method provides objective segmentations of GCL somas. The second difference between our results and previous studies was the distribution of the measured soma diameters. Previous literature [24, 157] has reported a bimodal distribution for the soma size at retinal eccentricities above  $6^\circ$ , with the peaks presumably corresponding to the mean diameters of parasol and midget cell populations. Although the distributions of our automatic diameter measurements for the  $12.75^\circ$  volumes did not appear bimodal for all subjects, a second smaller peak at higher diameter values was apparent for some subjects (S1 and S4 in Figure 22b and for randomly selected set of somas in Figure 23c). The difference between diameters measured from the manually and automatically segmented somas (Figure 23b) reflect that the automatic method yielded larger diameters for smaller GCL somas (diameters  $< 15 \mu\text{m}$ ) and smaller diameters for larger cells compared to the expert grader. These differences might ultimately render the two underlying peaks in the soma diameter distributions to be less distinguishable from each other over the set of all detected somas.

We showed the generalizability of our method to an unseen retinal location using IU's dataset. For this purpose, we used the AO-OCT volumes recorded at  $3.75^\circ$  and  $12.75^\circ$  retinal locations as the training data. When evaluated on the  $8.5^\circ$  AO-OCT volumes, the trained model achieved similar performance to the  $3.75^\circ$  and  $12.75^\circ$  dataset. As the two extreme locations involved in training encompassed the range of spatially varying GC

size, type, and density across much of the retina, we anticipate that the trained model would generalize to other untested retinal locations without additional training.

Although the performance of our method on the glaucoma dataset was lower than that on the healthy group, the expert performance on these data was lower as well. This reflects the inherent differences between the data from the two groups of subjects and the difficulty of identifying cells within glaucoma volumes. It is likely that with larger datasets, the performance of our method would improve.

To demonstrate the utility of our developed framework, we investigated the relationships between the automatically measured cellular-level characteristics of the GCL and its thickness values from AO-OCT and clinical OCT images and local functional measures from visual field test (TD and PD). Our results reflected larger GCL soma diameters in glaucoma subjects compared to healthy individuals. The structural analysis demonstrated a strong linear correlation between local GCL cell density and AO-OCT measured thickness across healthy and glaucoma subjects. Thickness values obtained from clinical OCT images, which were larger than values measured from AO-OCT images, did not exhibit strong relation to the local cell density. GCL cell density, AO-OCT measured thickness, and soma diameter (in log-scale) together increased the linear correlation with the functional measures in glaucoma subjects compared to when each measure was considered individually. As the population of glaucoma patients involved in this work was relatively small (five individuals) and the subjects varied in the stage of disease (early and moderate), future studies are needed to further investigate the

structure-function relationship throughout different stages of glaucoma. Our work paves the way towards these clinical studies.

In this work, we trained a CNN to localize somas in the AO-OCT volumes and then used the 3D watershed algorithm on the trained network's feature map to obtain individual soma masks. Future work could extend the current framework to additionally optimize the network's features for the instance segmentation task by incorporating regularization terms into the training loss function. Additionally, our work could be further extended by exploiting interactive instance segmentation techniques [167, 168] to correct errors from the automatically obtained segmentation masks with active guidance from an expert. Such an approach may also result in a framework robust to the inaccuracies in the initial user-provided training labels.

Despite the great potential of AO-OCT for early disease diagnosis and treatment outcome assessment, the lack of reliable automated soma quantification methods has impeded clinical translation. To our knowledge, this is the first automated GCL soma quantification method for AO-OCT volumes that achieved high detection performance and accurate soma diameter measurement, thus offering an attractive alternative to the costly and time-consuming manual marking process. We envision that our automated method would enable large-scale, multi-site clinical studies to further understand cellular-level pathological changes in retinal diseases.

## 5. Conclusion

The work presented in this dissertation described the development of frameworks for automatic analysis of neuronal signals and images. This included (1) quantification of the optimal resolution limits for temporally overlapping fluorescence signals of neural activity, (2) development of a fast and robust framework for automatic segmentation of active neurons, and (3) development of an automatic, weakly-supervised framework for volumetric segmentation of retinal ganglion cells. The performances of all developed frameworks were validated against either experimentally or manually obtained ground truth data and compared to other state-of-the-art methods, if applicable.

In chapters 2 and 3, frameworks for processing two-photon calcium imaging recordings of mice brain were developed. First, using a statistical approach, we showed that attaining resolution finer than the peak time for the genetically encoded calcium indicators GCaMP6s and GCaMP6f is possible. We further calculated the Cramer-Rao based lower bounds to obtain the limits of attainable precision in estimating the unknown AP-evoked fluorescence signal parameters for overlapping transients. In the next chapter, to overcome the bottleneck of fast neuron segmentation in the analysis workflow, we presented an automated, fast, and reliable active neuron segmentation framework with a 3D CNN named STNeuroNet. STNeuroNet was combined with intuitive pre- and post-processing steps for improved performance. Our framework sequentially processed calcium imaging recordings and could segment overlapping neurons, surpassed other

state-of-the-art methods on two separate two-photon microscopy datasets, was on par with expert graders, and could generalize to unseen data from other cortical layers.

In chapter 4, we shifted focus from segmenting neurons in two-photon calcium imaging data to segmenting retinal ganglion cells in AO-OCT images of human eyes. We designed and validated a new deep learning-based method with a 3D CNN at its core to automatically segment individual GCs. The elegant design of our overall framework in the context of weakly-supervised learning resulted in a high-fidelity software with performance on par with human experts which generalized to an unseen retinal location, between healthy and glaucoma subjects, and between imagers. Additionally, our method's soma segmentation masks agreed with manual segmentation masks and the mean and range of estimated soma diameter values were in the range reported in previous *ex vivo* and *in vivo* measurement studies. Through this technical advance, the processing time of AO-OCT volumes was significantly reduced from approximately 8 hours/volume to less than 3 minutes/volume.

In summary, the results of this dissertation provide opportunities for accurately exploring the dynamism of neuronal populations in real-time for neuroscience applications and will facilitate early diagnosis, accurate prognosis, and treatment monitoring of neurodegenerative retinal diseases like glaucoma for clinical use and research purposes.

## Appendix A: Calculation of the Fisher Information Matrix

Here we calculate the Fisher Information matrix. The elements of the Fisher information matrix for data with Poisson statistics are calculated as

$$\mathbf{I}_{Fij} = -\mathbb{E} \left\{ \frac{\partial^2 p(\mathbf{y}; \boldsymbol{\theta}_1)}{\partial \boldsymbol{\theta}_{1i} \partial \boldsymbol{\theta}_{1j}} \right\} = \sum_{k=1}^K \frac{1}{s_1(t_k; \boldsymbol{\theta}_1)} \frac{\partial s_1(t_k; \boldsymbol{\theta}_1)}{\partial \boldsymbol{\theta}_{1i}} \frac{\partial s_1(t_k; \boldsymbol{\theta}_1)}{\partial \boldsymbol{\theta}_{1j}}, \quad (\text{A.1})$$

in which  $\boldsymbol{\theta}_1$  contains the parameters defining the signal model. For the general case of unknown ISI ( $d$ ) and amplitudes ( $[\alpha, \beta]$ ),  $\boldsymbol{\theta}_1 = [d, \alpha, \beta]$ . As described in section 2.2.2.1 Spikes with known amplitudes, we are defining the two spike model such that the spike times ( $d_1$  and  $d_2$ ) are located around time  $t = 0$ .

$$s_1(t_k; \boldsymbol{\theta}_1) = \alpha F_0 h(t_k - d_1) + \beta F_0 h(t_k + d_2) + F_0. \quad (\text{A.2})$$

in which

$$d_1 = \frac{\beta}{\alpha + \beta} d \text{ and } d_2 = \frac{\alpha}{\alpha + \beta} d. \quad (\text{A.3})$$

The derivatives involved in calculating the elements of the Fisher Information matrix are derived below:

$$\frac{\partial s_1(t_k; d)}{\partial d} = \frac{-\alpha\beta}{\alpha + \beta} F_0 \left( \frac{\partial h}{\partial t} \Big|_{t=t_k-d_1} - \frac{\partial h}{\partial t} \Big|_{t=t_k+d_2} \right), \quad (\text{A.4})$$

$$\frac{\partial s_1(t_k; d, \alpha, \beta)}{\partial \alpha} = F_0 h \left( t_k - \frac{\beta}{\alpha + \beta} d \right) + \frac{\beta F_0 d}{(\alpha + \beta)^2} \left( \alpha \frac{\partial h}{\partial t} \Big|_{t=t_k-d_1} + \beta \frac{\partial h}{\partial t} \Big|_{t=t_k+d_2} \right), \quad (\text{A.5})$$

$$\frac{\partial s_1(t_k; d, \alpha, \beta)}{\partial \beta} = F_0 h \left( t_k + \frac{\alpha}{\alpha + \beta} d \right) - \frac{\alpha F_0 d}{(\alpha + \beta)^2} \left( \alpha \frac{\partial h}{\partial t} \Big|_{t=t_k-d_1} + \beta \frac{\partial h}{\partial t} \Big|_{t=t_k+d_2} \right), \quad (\text{A.6})$$

in which

$$\frac{\partial h}{\partial t} \Big|_{t=t_k-d_1} = \frac{-a}{\tau_d} e^{-\frac{t_k-d_1}{\tau_d}} u(t_k - d_1) + a \left( \frac{1}{\tau_d} + \frac{1}{\tau_{on}} \right) e^{-(t_k-d_1) \left( \frac{1}{\tau_d} + \frac{1}{\tau_{on}} \right)} u(t_k - d_1), \quad (\text{A.7})$$

and

$$\left. \frac{\partial h}{\partial t} \right|_{t=t_k+d_2} = \frac{-a}{\tau_d} e^{-\frac{t_k+d_2}{\tau_d}} u(t_k+d_2) + a \left( \frac{1}{\tau_d} + \frac{1}{\tau_{on}} \right) e^{-(t_k+d_2) \left( \frac{1}{\tau_d} + \frac{1}{\tau_{on}} \right)} u(t_k+d_2), \quad (\text{A.8})$$

## Appendix B: Other Algorithms for Automatic Neuron Segmentation from Two-photon Calcium imaging Videos.

### *B.1 CaImAn*

We used the available code at <https://github.com/flatironinstitute/CaImAn> to implement the algorithm of [18]. We selected the optimal parameter values for CaImAn Online and CaImAn Batch that resulted in the highest performance. Specifically, we performed a grid search over a range of values for the tuning parameters using leave-one-out cross-validation to quantify the performance on the ABO Layer 275  $\mu\text{m}$  data. We reported the performance scores on the ABO Layer 175  $\mu\text{m}$  test set and the Neurofinder test set using the best parameters determined by the ABO Layer 275  $\mu\text{m}$  data and the Neurofinder training data, respectively. The CaImAn toolbox includes two pre-trained CNNs for the analysis of calcium imaging data. One CNN is used during the processing pipeline of the CaImAn Online method, and the other is used as a post-processing step to reduce falsely detected masks. We have re-trained these two networks with the available data using the scripts provided by the authors.

When applying CaImAn Online to the ABO Layer 275  $\mu\text{m}$  data, we changed the expected half-size of neurons from 5 pixels to 10 pixels (3.9  $\mu\text{m}$  to 7.8  $\mu\text{m}$ ) and selected the number of components during the initialization phase from [2, 10, 50, 150] and the number of frames for initialization from [100, 200, 300]. We selected the minimum signal-to-noise ratio (SNR) for accepting new components from [2, 4, 6, 8], the maximum number

of neurons added per frame from [5, 10, 25, 50], the threshold of the trained classifiers for adding new components during the online processing from [0.5, 0.7, 0.8, 0.9, 0.95], and the threshold for eliminating false positives from 0 to 0.5 with step size of 0.01. When applying CaImAn Online to the Neurofinder dataset, we set the expected half-size of neurons, the initial batch size and the number of initialization components to the values used by [18]. We selected the minimum acceptable SNR from [2, 2.5, 4, 6] and the maximum number of added neurons per frame from [5, 10, 20] while changing the classifier thresholds for adding new components and eliminating components from [0.5, 0.75, 0.8] and from 0 to 0.5 with step size of 0.1, respectively.

For CaImAn Batch on both ABO and Neurofinder datasets, we used the optimal half-size of neurons found from the CaImAn Online results. We set the patches to be 100×100 pixels with 10 pixels overlap between patches. We set the number of components per patch to 40, twice the maximum average number of neurons per 100×100 pixels area from the GT set, to avoid low recall. We selected the spatial correlation threshold from [0.75, 0.80, 0.85], the upper and lower thresholds for the CNN classifier from [0.8, 0.9, 0.95, 0.98] and 0 to 0.5 with step size of 0.1, respectively. We selected the minimum SNR for the ABO dataset from 4 to 10 with increment of 2, and for the Neurofinder dataset from [1.8, 2, 2.5, 3]. We used the optimal values that yielded the highest mean  $F_1$  score across the training set to quantify the final performances. As in [18], we binarized each real-valued detected mask by using 0.2 times the maximum value of the mask as the threshold.

## ***B.2 Suite2p***

We used the code provided by [20] available online at <https://github.com/cortex-lab/Suite2P>. Through leave-one-out cross-validation, we quantified the performance of Suite2p on the ABO Layer 275  $\mu\text{m}$  dataset. We used all of the ABO Layer 275  $\mu\text{m}$  data and Neurofinder training data to quantify the performance on the ABO Layer 175  $\mu\text{m}$  test set and the Neurofinder test set, respectively. For both the ABO and Neurofinder datasets, we varied the diameter of neurons from 7.8  $\mu\text{m}$  to 15.6  $\mu\text{m}$  with step size of 3.9  $\mu\text{m}$ , the number of singular value decomposition (SVD) components from 200 to 800 with step size of 100, number of frames for SVD from 1000 to 4000 in steps of 1000. We selected the probability threshold of their ROI classifier from 0 to 0.5 with step size of 0.1, and the minimum and maximum acceptable sizes from 15 to 120  $\mu\text{m}^2$  and 100 to 845  $\mu\text{m}^2$  in increments of 18  $\mu\text{m}^2$  and 426  $\mu\text{m}^2$ , respectively. We kept all other parameters at the default values set by the authors of [20]. For each data, we ran the Suite2p procedure until the number of detected neurons did not change, or until we reached a maximum of one hundred iterations. We also trained their ROI classifier on the training videos by manually curating the results that yielded the largest number of detected neurons. For each validation iteration, we used the best combination of parameters that yielded the highest mean  $F_1$  score on the training data for the test data to report the final performance scores of Suite2p.

### ***B.3 HNCcorr***

HNCcorr is a graph-cut based method that processes the correlation image. We used the code provided by [64] at <https://github.com/hochbaumGroup/HNCcorr>. Like other methods, we performed leave-one-out cross-validation to quantify HNCcorr's performance on the ABO Layer 275  $\mu\text{m}$  data, and used all of the ABO Layer 275  $\mu\text{m}$  data and the Neurofinder training data to quantify the performance on the ABO Layer 175  $\mu\text{m}$  test set and the Neurofinder test set, respectively. For the ABO dataset, we set the segmentation window size to 37 pixels (28.9  $\mu\text{m}$ ) and the average neuron size to 107.5  $\mu\text{m}^2$ . We selected the percentage of seeds from 0.1 to 0.7 with step size of 0.1, the seed size from 1 $\times$ 1 pixel to 5 $\times$ 5 pixels, and the minimum and maximum acceptable sizes from 35 to 60  $\mu\text{m}^2$  and 122 to 243  $\mu\text{m}^2$  with step size of 6  $\mu\text{m}^2$  and 30  $\mu\text{m}^2$ , respectively. For the Neurofinder dataset, based on the parameters reported in [64], we set the segmentation window size to 41 pixels, the percentage of seeds to 0.4, and the average neuron sizes to the values reported in Supplementary Table 3 of [64]. In accordance with the values used in [64], we changed the seed size from 1 $\times$ 1 pixel to 5 $\times$ 5 pixels, and the minimum and maximum acceptable sizes from 30 to 50 pixels and 200 to 800 pixels with step size of 10 and 100 pixels, respectively. We used the combination of parameters that yielded the highest mean  $F_1$  score on the training data for the test data to report the final performance scores of HNCcorr.

## ***B.4 UNet2DS***

We used the code provided by [67] at <https://github.com/alexklibisz/deep-calcium> to train and test the UNet2DS network on the ABO dataset. This CNN is based on the popular UNet [70] and uses the mean image of the data to segment neurons. We performed leave-one-out cross-validation to quantify the performance of this network on the ABO Layer 275  $\mu\text{m}$  data, and used the ABO Layer 275  $\mu\text{m}$  data and the Neurofinder training data to quantify the performance on the ABO Layer 175  $\mu\text{m}$  test set and Neurofinder test set, respectively. Using the same training procedure outlined by [67], we trained UNet2DS for 50 epochs with 100 training iterations in each epoch using sixteen randomly cropped 128 $\times$ 128 pixels regions from the mean image, utilizing the dice-loss and the Adam optimizer. In accordance with [67], we tracked the  $F_1$  score on a validation video selected from the training set to ensure the network was not overfitting. For the Neurofinder data, we used the exact scripts provided by [67] to train on the six training videos. At inference time, we averaged the predictions from eight rotations and reflections of the full spatial-extent of the test image to make the final prediction. We used the combination of parameters that yielded the highest mean  $F_1$  score on the training data for the test data to report the final performance scores.

## References

- [1] L. Grosenick, J. H. Marshel, and K. Deisseroth, "Closed-loop and activity-guided optogenetic control," *Neuron*, vol. 86, no. 1, pp. 106-139, 2015.
- [2] A. M. Packer, L. E. Russell, H. W. Dalglish, and M. Häusser, "Simultaneous all-optical manipulation and recording of neural circuit activity with cellular resolution in vivo," *Nature methods*, vol. 12, no. 2, p. 140, 2015.
- [3] J. P. Rickgauer, K. Deisseroth, and D. W. Tank, "Simultaneous cellular-resolution optical perturbation and imaging of place cell firing fields," *Nature neuroscience*, vol. 17, no. 12, pp. 1816-1824, 2014.
- [4] T. Deneux *et al.*, "Accurate spike estimation from noisy calcium signals for ultrafast three-dimensional imaging of large neuronal populations in vivo," *Nature communications*, vol. 7, p. 12190, 2016.
- [5] B. F. Grewe, D. Langer, H. Kasper, B. M. Kampa, and F. Helmchen, "High-speed in vivo calcium imaging reveals neuronal network activity with near-millisecond precision," *Nat Methods*, vol. 7, no. 5, pp. 399-405, May 2010.
- [6] J. Onativia, S. R. Schultz, and P. L. Dragotti, "A finite rate of innovation algorithm for fast and accurate spike detection from two-photon calcium imaging," *Journal of neural engineering*, vol. 10, no. 4, p. 046017, 2013.
- [7] E. A. Pnevmatikakis, J. Merel, A. Pakman, and L. Paninski, "Bayesian spike inference from calcium imaging data," in *Signals, Systems and Computers, 2013 Asilomar Conference on*, 2013, pp. 349-353: IEEE.
- [8] A. F. Szymanska, C. Kobayashi, H. Norimoto, T. Ishikawa, Y. Ikegaya, and Z. Nenadic, "Accurate detection of low signal-to-noise ratio neuronal calcium transient waves using a matched filter," *Journal of neuroscience methods*, vol. 259, pp. 1-12, 2016.
- [9] L. Theis *et al.*, "Benchmarking Spike Rate Inference in Population Calcium Imaging," *Neuron*, vol. 90, no. 3, pp. 471-82, May 4 2016.

- [10] J. T. Vogelstein *et al.*, "Fast nonnegative deconvolution for spike train inference from population calcium imaging," *J Neurophysiol*, vol. 104, no. 6, pp. 3691-704, Dec 2010.
- [11] J. T. Vogelstein, B. O. Watson, A. M. Packer, R. Yuste, B. Jedynek, and L. Paninski, "Spike inference from calcium imaging using sequential Monte Carlo methods," *Biophys J*, vol. 97, no. 2, pp. 636-55, Jul 22 2009.
- [12] B. A. Wilt, J. E. Fitzgerald, and M. J. Schnitzer, "Photon shot noise limits on optical detection of neuronal spikes and estimation of spike timing," *Biophys J*, vol. 104, no. 1, pp. 51-62, Jan 8 2013.
- [13] M. Pachitariu, A. M. Packer, N. Pettit, H. Dalgleish, M. Hausser, and M. Sahani, "Extracting regions of interest from biological images with convolutional sparse block coding," in *Advances in Neural Information Processing Systems*, 2013, pp. 1745-1753.
- [14] E. A. Mukamel, A. Nimmerjahn, and M. J. Schnitzer, "Automated analysis of cellular signals from large-scale calcium imaging data," *Neuron*, vol. 63, no. 6, pp. 747-60, Sep 24 2009.
- [15] E. A. Pnevmatikakis *et al.*, "A structured matrix factorization framework for large scale calcium imaging data analysis," *ArXiv preprint*, 2014.
- [16] E. A. Pnevmatikakis *et al.*, "Simultaneous Denoising, Deconvolution, and Demixing of Calcium Imaging Data," *Neuron*, vol. 89, no. 2, pp. 285-99, Jan 20 2016.
- [17] F. D. Andilla and F. A. Hamprecht, "Sparse space-time deconvolution for calcium image analysis," in *Advances in Neural Information Processing Systems*, 2014, pp. 64-72.
- [18] A. Giovannucci *et al.*, "CaImAn: An open source tool for scalable Calcium Imaging data Analysis," *eLife*, vol. 8, p. e38173, 2019.

- [19] A. Giovannucci *et al.*, "OnACID: Online Analysis of Calcium Imaging Data in Real Time," in *Advances in Neural Information Processing Systems*, 2017, pp. 2378-2388.
- [20] M. Pachitariu *et al.*, "Suite2p: beyond 10,000 neurons with standard two-photon microscopy," *BioRxiv*, p. 061507, 2017.
- [21] S. Kingman. (2004). *Glaucoma is second leading cause of blindness globally*. Available: <https://www.who.int/bulletin/volumes/82/11/feature1104/en/>
- [22] R. N. Weinreb and P. T. Khaw, "Primary open-angle glaucoma," *The Lancet*, vol. 363, no. 9422, pp. 1711-1720, 2004.
- [23] R. Werkmeister, A. P. Cherecheanu, G. Garhofer, D. Schmidl, and L. Schmetterer, "Imaging of retinal ganglion cells in glaucoma: pitfalls and challenges," *Cell and tissue research*, vol. 353, no. 2, pp. 261-268, 2013.
- [24] Z. Liu, K. Kurokawa, F. Zhang, J. J. Lee, and D. T. Miller, "Imaging and quantifying ganglion cells and other transparent neurons in the living human retina," *Proceedings of the National Academy of Sciences*, vol. 114, no. 48, pp. 12803-12808, 2017.
- [25] Z. Liu, J. Tam, O. Saeedi, and D. X. Hammer, "Trans-retinal cellular imaging with multimodal adaptive optics," *Biomedical optics express*, vol. 9, no. 9, pp. 4246-4262, 2018.
- [26] S. Soltanian-Zadeh, Y. Gong, and S. Farsiu, "Information-theoretic approach and fundamental limits of resolving two closely timed neuronal spikes in mouse brain calcium imaging," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 11, pp. 2428-2439, 2018.
- [27] L. Tian, S. A. Hires, and L. L. Looger, "Imaging neuronal activity with genetically encoded calcium indicators," *Cold Spring Harbor Protocols*, vol. 2012, no. 6, p. pdb.top069609, 2012.

- [28] J. L. Chen, M. L. Andermann, T. Keck, N.-L. Xu, and Y. Ziv, "Imaging neuronal populations in behaving rodents: paradigms for studying neural circuits underlying behavior in the mammalian cortex," *Journal of Neuroscience*, vol. 33, no. 45, pp. 17631-17640, 2013.
- [29] J. T. Vogelstein *et al.*, "Fast nonnegative deconvolution for spike train inference from population calcium imaging," *Journal of neurophysiology*, vol. 104, no. 6, pp. 3691-3704, 2010.
- [30] L. Theis *et al.*, "Benchmarking spike rate inference in population calcium imaging," *Neuron*, vol. 90, no. 3, pp. 471-482, 2016.
- [31] E. A. Pnevmatikakis, J. Merel, A. Pakman, and L. Paninski, "Bayesian spike inference from calcium imaging data," in *2013 Asilomar Conference on Signals, Systems and Computers*, 2013, pp. 349-353: IEEE.
- [32] E. L. Dyer, M. F. Duarte, D. H. Johnson, and R. G. Baraniuk, "Recovering spikes from noisy neuronal calcium signals via structured sparse approximation," in *International Conference on Latent Variable Analysis and Signal Separation*, 2010, pp. 604-611: Springer.
- [33] T. Quan, X. Lv, X. Liu, and S. Zeng, "Reconstruction of burst activity from calcium imaging of neuronal population via Lq minimization and interval screening," *Biomedical optics express*, vol. 7, no. 6, pp. 2103-2117, 2016.
- [34] E. L. Dyer, C. Studer, J. T. Robinson, and R. G. Baraniuk, "A robust and efficient method to recover neural events from noisy and corrupted data," in *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*, 2013, pp. 593-596: IEEE.
- [35] J. Friedrich, P. Zhou, and L. Paninski, "Fast online deconvolution of calcium imaging data," *PLoS computational biology*, vol. 13, no. 3, p. e1005423, 2017.
- [36] B. F. Grewe, D. Langer, H. Kasper, B. M. Kampa, and F. Helmchen, "High-speed in vivo calcium imaging reveals neuronal network activity with near-millisecond precision," *Nature methods*, vol. 7, no. 5, p. 399, 2010.

- [37] B. A. Wilt, J. E. Fitzgerald, and M. J. Schnitzer, "Photon shot noise limits on optical detection of neuronal spikes and estimation of spike timing," *Biophysical journal*, vol. 104, no. 1, pp. 51-62, 2013.
- [38] S. Reynolds, J. Onativia, C. S. Copeland, S. R. Schultz, and P. L. Dragotti, "Spike detection using FRI methods and protein calcium sensors: performance analysis and comparisons," in *2015 International Conference on Sampling Theory and Applications (SampTA)*, 2015, pp. 533-537: IEEE.
- [39] S. Ram, E. S. Ward, and R. J. Ober, "Beyond Rayleigh's criterion: a resolution measure with application to single-molecule microscopy," *Proceedings of the National Academy of Sciences*, vol. 103, no. 12, pp. 4457-4462, 2006.
- [40] S. Ram, E. S. Ward, and R. J. Ober, "A stochastic analysis of distance estimation approaches in single molecule microscopy: quantifying the resolution limits of photon-limited imaging systems," *Multidimensional systems and signal processing*, vol. 24, no. 3, pp. 503-542, 2013.
- [41] M. Shahram and P. Milanfar, "Imaging below the diffraction limit: a statistical analysis," *IEEE Transactions on image processing*, vol. 13, no. 5, pp. 677-689, 2004.
- [42] S. Van Aert, D. Van Dyck, and J. Arnold, "Resolution of coherent and incoherent imaging systems reconsidered—Classical criteria and a statistical alternative," *Optics express*, vol. 14, no. 9, pp. 3830-3839, 2006.
- [43] S. Farsiu *et al.*, "Statistical detection and imaging of objects hidden in turbid media using ballistic photons," *Applied optics*, vol. 46, no. 23, pp. 5805-5822, 2007.
- [44] F. Helmchen, K. Imoto, and B. Sakmann, "Ca<sup>2+</sup> buffering and action potential-evoked Ca<sup>2+</sup> signaling in dendrites of pyramidal neurons," *Biophysical journal*, vol. 70, no. 2, pp. 1069-1081, 1996.
- [45] T.-W. Chen *et al.*, "Ultrasensitive fluorescent proteins for imaging neuronal activity," *Nature*, vol. 499, no. 7458, pp. 295-300, 2013.
- [46] S. M. Kay, *Fundamentals of statistical signal processing*. Prentice Hall PTR, 1993.

- [47] H. L. Van Trees and K. L. Bell, *Detection Estimation and Modulation Theory, Detection, Estimation, and Filtering Theory, Volume 1 (2)*. Somerset, US: Wiley, 2013.
- [48] P. R. Bevington and D. K. Robinson, "Data reduction and error analysis," *McGraw-Hill*, 2003.
- [49] D. A. Turton, G. D. Reid, and G. S. Beddard, "Accurate analysis of fluorescence decays from single molecules in photon counting experiments," *Analytical chemistry*, vol. 75, no. 16, pp. 4182-4187, 2003.
- [50] K. P. Burnham and D. R. Anderson, "Multimodel inference understanding AIC and BIC in model selection," *Sociological methods & research*, vol. 33, no. 2, pp. 261-304, 2004.
- [51] S. M. Kay, "Fundamentals of statistical signal processing, volume I: estimation theory," 1993.
- [52] C. Forbes, M. Evans, N. Hastings, and B. Peacock, *Statistical distributions*. Hoboken, New Jersey: John Wiley & Sons, 2011.
- [53] P. Dayan and L. F. Abbott, *Theoretical neuroscience*. Cambridge, MA: MIT Press, 2001.
- [54] C. J. Clopper and E. S. Pearson, "The use of confidence or fiducial limits illustrated in the case of the binomial," *Biometrika*, pp. 404-413, 1934.
- [55] P. J. Huber, "The behavior of maximum likelihood estimates under nonstandard conditions," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 1967, vol. 1, no. 1, pp. 221-233.
- [56] S. Ram, E. S. Ward, and R. J. Ober, "Beyond Rayleigh's criterion: a resolution measure with application to single-molecule microscopy," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 12, pp. 4457-4462, 2006.

- [57] R. Brette and A. Destexhe, *Handbook of neural activity measurement*. Cambridge University Press, 2012.
- [58] V. Rahmati, K. Kirmse, D. Marković, K. Holthoff, and S. J. Kiebel, "Inferring neuronal dynamics from calcium imaging data using biophysical models and Bayesian inference," *PLoS computational biology*, vol. 12, no. 2, p. e1004736, 2016.
- [59] S. Soltanian-Zadeh, K. Sahingur, S. Blau, Y. Gong, and S. Farsiu, "Fast and robust active neuron segmentation in two-photon calcium imaging using spatiotemporal deep learning," *Proceedings of the National Academy of Sciences*, vol. 116, no. 17, pp. 8554-8563, 2019.
- [60] J. Guan *et al.*, "NeuroSeg: automated cell detection and segmentation for in vivo two-photon Ca<sup>2+</sup> imaging data," *Brain Structure and Function*, vol. 223, no. 1, pp. 519-533, 2018.
- [61] P. Kaifosh, J. D. Zaremba, N. B. Danielson, and A. Losonczy, "SIMA: Python software for analysis of dynamic fluorescence imaging data," *Frontiers in neuroinformatics*, vol. 8, p. 80, 2014.
- [62] R. Maruyama *et al.*, "Detecting cells using non-negative matrix factorization on calcium imaging data," *Neural Networks*, vol. 55, pp. 11-19, 2014.
- [63] S. Reynolds, T. Abrahamsson, R. Schuck, P. J. Sjöström, S. R. Schultz, and P. L. Dragotti, "ABLE: An Activity-Based Level Set Segmentation Algorithm for Two-Photon Calcium Imaging Data," *eNeuro*, vol. 4, no. 5, pp. ENEURO.0012-17.2017, 2017.
- [64] Q. Spaen, D. S. Hochbaum, and R. Asín-Achá, "HNCcorr: A Novel Combinatorial Approach for Cell Identification in Calcium-Imaging Movies," *arXiv preprint arXiv:1703.01999*, 2017.
- [65] P. Zhou *et al.*, "Efficient and accurate extraction of in vivo calcium signals from microendoscopic video data," *eLife*, vol. 7, p. e28728, 2018.

- [66] N. Apthorpe *et al.*, "Automatic neuron detection in calcium imaging data using convolutional networks," in *Advances in Neural Information Processing Systems*, 2016, pp. 3270-3278.
- [67] A. Klibisz, D. Rose, M. Eicholtz, J. Blundon, and S. Zakharenko, "Fast, Simple Calcium Imaging Segmentation with Fully Convolutional Networks," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*: Springer, 2017, pp. 285-293.
- [68] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436-444, 05/27/online 2015.
- [69] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431-3440.
- [70] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 234-241: Springer.
- [71] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Computer Vision (ICCV), 2015 IEEE International Conference on*, 2015, pp. 4489-4497: IEEE.
- [72] G. Varol, I. Laptev, and C. Schmid, "Long-term temporal convolutions for action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 6, pp. 1510-1517, 2017.
- [73] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: learning dense volumetric segmentation from sparse annotation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2016, pp. 424-432: Springer.
- [74] K. Kamnitsas *et al.*, "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation," *Medical image analysis*, vol. 36, pp. 61-78, 2017.

- [75] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *3D Vision (3DV), 2016 Fourth International Conference on*, 2016, pp. 565-571: IEEE.
- [76] S. E. de Vries *et al.*, "A large-scale, standardized physiological survey reveals higher order coding throughout the mouse visual cortex," *bioRxiv*, p. 359513, 2018.
- [77] A. v. Oppenheim, R. Schafer, and T. Stockham, "Nonlinear filtering of multiplied and convolved signals," *IEEE transactions on audio and electroacoustics*, vol. 16, no. 3, pp. 437-466, 1968.
- [78] E. Gibson *et al.*, "Automatic multi-organ segmentation on abdominal CT with dense v-networks," *IEEE Transactions on Medical Imaging*, vol. 37, no. 8, pp. 1822-1834, 2018.
- [79] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," in *International Conference on Machine Learning*, 2015, pp. 448-456.
- [80] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929-1958, 2014.
- [81] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [82] F. Meyer, "Topographic distance and watershed lines," *Signal processing*, vol. 38, no. 1, pp. 113-125, 1994.
- [83] S. Soltanian-Zadeh, Y. Gong, and S. Farsiu, "Information-Theoretic Approach and Fundamental Limits of Resolving Two Closely-Timed Neuronal Spikes in Mouse Brain Calcium Imaging," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 11, pp. 2428-2439, 08 March 2018 2018.

- [84] C. M. Niell and M. P. Stryker, "Modulation of visual responses by behavioral state in mouse visual cortex," *Neuron*, vol. 65, no. 4, pp. 472-479, 2010.
- [85] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural computation*, vol. 7, no. 6, pp. 1129-1159, 1995.
- [86] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," in *Advances in neural information processing systems*, 2014, pp. 3320-3328.
- [87] D. Rolnick, A. Veit, S. Belongie, and N. Shavit, "Deep learning is robust to massive label noise," *arXiv preprint arXiv:1705.10694*, 2017.
- [88] S. Sukhbaatar, J. Bruna, M. Paluri, L. Bourdev, and R. Fergus, "Training convolutional networks with noisy labels," *arXiv preprint arXiv:1406.2080*, 2014.
- [89] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," *arXiv preprint arXiv:1611.03530*, 2016.
- [90] E. A. Pnevmatikakis and A. Giovannucci, "NoRMCorre: An online algorithm for piecewise rigid motion correction of calcium imaging data," *Journal of neuroscience methods*, vol. 291, pp. 83-94, 2017.
- [91] S. Soltanian-Zadeh, K. Kurokawa, Z. Liu, D. X. Hammer, D. T. Miller, and S. Farsiu, "Fully automatic quantification of individual ganglion cells from AO-OCT volumes via weakly supervised learning," in *Ophthalmic Technologies XXX*, 2020, vol. 11218, p. 112180Q: International Society for Optics and Photonics.
- [92] S. Soltanian-Zadeh *et al.*, "Automatic cellular level differentiation of glaucomatous and healthy eyes via deep learning-based adaptive optics OCT analysis," *Investigative Ophthalmology & Visual Science*, vol. 61, no. 7, pp. 877-877, 2020.

- [93] S. Kingman, "Glaucoma is second leading cause of blindness globally," *Bulletin of the World Health Organization*, vol. 82, pp. 887-888, 2004.
- [94] A. Tafreshi *et al.*, "Visual function-specific perimetry to identify glaucomatous visual loss using three different definitions of visual field abnormality," *Investigative ophthalmology & visual science*, vol. 50, no. 3, pp. 1234-1240, 2009.
- [95] D. B. Henson, S. Chaudry, P. H. Artes, E. B. Faragher, and A. Ansons, "Response variability in the visual field: comparison of optic neuritis, glaucoma, ocular hypertension, and normal eyes," *Investigative Ophthalmology & Visual Science*, vol. 41, no. 2, pp. 417-421, 2000.
- [96] A. Heijl, A. Lindgren, and G. Lindgren, "Test-retest variability in glaucomatous visual fields," *American journal of ophthalmology*, vol. 108, no. 2, pp. 130-135, 1989.
- [97] R. S. Harwerth, M. Crawford, L. J. Frishman, S. Viswanathan, E. L. Smith Iii, and L. Carter-Dawson, "Visual field defects and neural losses from experimental glaucoma," *Progress in retinal and eye research*, vol. 21, no. 1, pp. 91-125, 2002.
- [98] A. J. Tatham and F. A. Medeiros, "Detecting structural progression in glaucoma with optical coherence tomography," *Ophthalmology*, vol. 124, no. 12, pp. S57-S65, 2017.
- [99] Z. M. Dong, G. Wollstein, and J. S. Schuman, "Clinical utility of optical coherence tomography in glaucoma," *Investigative ophthalmology & visual science*, vol. 57, no. 9, pp. OCT556-OCT567, 2016.
- [100] T. M. Kuang, C. Zhang, L. M. Zangwill, R. N. Weinreb, and F. A. Medeiros, "Estimating lead time gained by optical coherence tomography in detecting glaucoma before development of visual field defects," *Ophthalmology*, vol. 122, no. 10, pp. 2002-2009, 2015.
- [101] T. E. Ogden, "Nerve fiber layer of the primate retina: morphometric analysis," *Investigative ophthalmology & visual science*, vol. 25, no. 1, pp. 19-29, 1984.

- [102] C. Bowd, L. M. Zangwill, R. N. Weinreb, F. A. Medeiros, and A. Belghith, "Estimating optical coherence tomography structural measurement floors to improve detection of progression in advanced glaucoma," *American journal of ophthalmology*, vol. 175, pp. 37-44, 2017.
- [103] K. Banister *et al.*, "Can automated imaging for optic disc and retinal nerve fiber layer analysis aid glaucoma detection?," *Ophthalmology*, vol. 123, no. 5, pp. 930-938, 2016.
- [104] J.-C. Mwanza *et al.*, "Retinal nerve fibre layer thickness floor and corresponding functional loss in glaucoma," *British Journal of Ophthalmology*, vol. 99, no. 6, pp. 732-737, 2015.
- [105] F. A. Medeiros, A. A. Jammal, and A. C. Thompson, "From machine to machine: an OCT-trained deep learning algorithm for objective quantification of glaucomatous damage in fundus photographs," *Ophthalmology*, vol. 126, no. 4, pp. 513-521, 2019.
- [106] H. Fu *et al.*, "Disc-aware ensemble network for glaucoma screening from fundus image," *IEEE transactions on medical imaging*, vol. 37, no. 11, pp. 2493-2501, 2018.
- [107] M. Christopher *et al.*, "Deep learning approaches predict glaucomatous visual field damage from OCT optic nerve head En face images and retinal nerve fiber layer thickness maps," *Ophthalmology*, vol. 127, no. 3, pp. 346-356, 2020.
- [108] Y. M. George, B. Antony, H. Ishikawa, G. Wollstein, J. S. Schuman, and R. Garnavi, "Attention-guided 3D-CNN Framework for Glaucoma Detection and Structural-Functional Association using Volumetric Images," *IEEE Journal of Biomedical and Health Informatics*, 2020.
- [109] E. M. Wells-Gray, S. S. Choi, M. Slabaugh, P. Weber, and N. Doble, "Inner retinal changes in primary open-angle glaucoma revealed through adaptive optics-optical coherence tomography," *Journal of glaucoma*, vol. 27, no. 11, pp. 1025-1028, 2018.

- [110] K. Kurokawa, J. A. Crowell, F. Zhang, and D. T. Miller, "Suite of methods for assessing inner retinal temporal dynamics across spatial and temporal scales in the living human eye," *Neurophotonics*, vol. 7, no. 1, p. 015013, 2020.
- [111] E. A. Rossi *et al.*, "Imaging individual neurons in the retinal ganglion cell layer of the living eye," *Proceedings of the National Academy of Sciences*, vol. 114, no. 3, pp. 586-591, 2017.
- [112] S. J. Chiu, X. T. Li, P. Nicholas, C. A. Toth, J. A. Izatt, and S. Farsiu, "Automatic segmentation of seven retinal layers in SDOCT images congruent with expert manual segmentation," *Optics express*, vol. 18, no. 18, pp. 19413-19428, 2010.
- [113] L. Fang, D. Cunefare, C. Wang, R. H. Guymer, S. Li, and S. Farsiu, "Automatic segmentation of nine retinal layer boundaries in OCT images of non-exudative AMD patients using deep learning and graph search," *Biomedical optics express*, vol. 8, no. 5, pp. 2732-2744, 2017.
- [114] J. Kugelman, D. Alonso-Caneiro, S. A. Read, S. J. Vincent, and M. J. Collins, "Automatic segmentation of OCT retinal boundaries using recurrent neural networks and graph search," *Biomedical optics express*, vol. 9, no. 11, pp. 5759-5777, 2018.
- [115] A. G. Roy *et al.*, "ReLayNet: retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks," *Biomedical optics express*, vol. 8, no. 8, pp. 3627-3642, 2017.
- [116] R. Kafieh, H. Rabbani, M. D. Abramoff, and M. Sonka, "Intra-retinal layer segmentation of 3D optical coherence tomography using coarse grained diffusion map," *Medical image analysis*, vol. 17, no. 8, pp. 907-928, 2013.
- [117] H. Fu, Y. Xu, S. Lin, D. W. K. Wong, and J. Liu, "Deepvessel: Retinal vessel segmentation via deep learning and conditional random field," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2016, pp. 132-139: Springer.

- [118] Q. Li, B. Feng, L. Xie, P. Liang, H. Zhang, and T. Wang, "A Cross-Modality Learning Approach for Vessel Segmentation in Retinal Images," *IEEE Trans. Med. Imaging*, vol. 35, no. 1, pp. 109-118, 2016.
- [119] P. Liskowski and K. Krawiec, "Segmenting retinal blood vessels with deep neural networks," *IEEE transactions on medical imaging*, vol. 35, no. 11, pp. 2369-2380, 2016.
- [120] T. B. Sekou, M. Hidane, J. Olivier, and H. Cardot, "Retinal Blood Vessel Segmentation Using a Fully Convolutional Network–Transfer Learning from Patch-to Image-Level," in *International Workshop on Machine Learning in Medical Imaging*, 2018, pp. 170-178: Springer.
- [121] Z. Yan, X. Yang, and K.-T. T. Cheng, "Joint Segment-level and Pixel-wise Losses for Deep Learning based Retinal Vessel Segmentation," *IEEE Transactions on Biomedical Engineering*, 2018.
- [122] H. Zhao, H. Li, S. Maurer-Stroh, Y. Guo, Q. Deng, and L. Cheng, "Supervised Segmentation of Un-annotated Retinal Fundus Images by Synthesis," *IEEE transactions on medical imaging*, 2018.
- [123] R. Estrada, C. Tomasi, M. T. Cabrera, D. K. Wallace, S. F. Freedman, and S. Farsiu, "Exploratory Dijkstra forest based automatic vessel segmentation: applications in video indirect ophthalmoscopy (VIO)," *Biomedical optics express*, vol. 3, no. 2, pp. 327-339, 2012.
- [124] D. Cunefare *et al.*, "Deep learning based detection of cone photoreceptors with multimodal adaptive optics scanning light ophthalmoscope images of achromatopsia," *Biomedical optics express*, vol. 9, no. 8, pp. 3740-3756, 2018.
- [125] D. Cunefare *et al.*, "Automatic detection of cone photoreceptors in split detector adaptive optics scanning light ophthalmoscope images," *Biomedical optics express*, vol. 7, no. 5, pp. 2036-2050, 2016.
- [126] S. J. Chiu, C. A. Toth, C. B. Rickman, J. A. Izatt, and S. Farsiu, "Automatic segmentation of closed-contour features in ophthalmic images using graph

theory and dynamic programming," *Biomedical optics express*, vol. 3, no. 5, pp. 1127-1140, 2012.

- [127] S. J. Chiu *et al.*, "Automatic cone photoreceptor segmentation using graph theory and dynamic programming," *Biomedical optics express*, vol. 4, no. 6, pp. 924-937, 2013.
- [128] B. Davidson *et al.*, "Automatic cone photoreceptor localisation in healthy and Stargardt afflicted retinas using deep learning," *Scientific reports*, vol. 8, no. 1, p. 7911, 2018.
- [129] M. Heisler *et al.*, "Automated identification of cone photoreceptors in adaptive optics optical coherence tomography images using transfer learning," *Biomedical optics express*, vol. 9, no. 11, pp. 5353-5367, 2018.
- [130] J. Loo, L. Fang, D. Cunefare, G. J. Jaffe, and S. Farsiu, "Deep longitudinal transfer learning-based automatic segmentation of photoreceptor ellipsoid zone defects on optical coherence tomography images of macular telangiectasia type 2," *Biomedical Optics Express*, vol. 9, no. 6, pp. 2681-2698, 2018.
- [131] F. G. Venhuizen *et al.*, "Deep learning approach for the detection and quantification of intraretinal cystoid fluid in multivendor optical coherence tomography," *Biomedical optics express*, vol. 9, no. 4, pp. 1545-1569, 2018.
- [132] J. De Fauw *et al.*, "Clinically applicable deep learning for diagnosis and referral in retinal disease," *Nature medicine*, vol. 24, no. 9, p. 1342, 2018.
- [133] T. Falk *et al.*, "U-Net: deep learning for cell counting, detection, and morphometry," *Nature methods*, vol. 16, no. 1, p. 67, 2019.
- [134] H. Chen, Q. Dou, L. Yu, J. Qin, and P.-A. Heng, "VoxResNet: Deep voxelwise residual networks for brain segmentation from 3D MR images," *NeuroImage*, vol. 170, pp. 446-455, 2018.

- [135] O. Oktay *et al.*, "Anatomically constrained neural networks (ACNNs): application to cardiac image enhancement and segmentation," *IEEE transactions on medical imaging*, vol. 37, no. 2, pp. 384-395, 2018.
- [136] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes," *IEEE transactions on medical imaging*, vol. 37, no. 12, pp. 2663-2674, 2018.
- [137] Y. Zhou, Y. Zhu, Q. Ye, Q. Qiu, and J. Jiao, "Weakly supervised instance segmentation using class peak response," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3791-3800.
- [138] K.-J. Hsu, Y.-Y. Lin, and Y.-Y. Chuang, "Deepco3: Deep instance co-segmentation by co-peak search and cosaliency detection," in *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [139] D. Lin, J. Dai, J. Jia, K. He, and J. Sun, "Scribblesup: Scribble-supervised convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3159-3167.
- [140] M. Tang, A. Djelouah, F. Perazzi, Y. Boykov, and C. Schroers, "Normalized cut loss for weakly-supervised CNN segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1818-1827.
- [141] M. Tang, F. Perazzi, A. Djelouah, I. Ben Ayed, C. Schroers, and Y. Boykov, "On regularized losses for weakly-supervised cnn segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 507-522.
- [142] A. Khoreva, R. Benenson, J. Hosang, M. Hein, and B. Schiele, "Simple does it: Weakly supervised instance and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 876-885.
- [143] J. Dai, K. He, and J. Sun, "Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1635-1643.

- [144] G. Papandreou, L.-C. Chen, K. P. Murphy, and A. L. Yuille, "Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1742-1750.
- [145] M. Rajchl *et al.*, "Deepcut: Object segmentation from bounding box annotations using convolutional neural networks," *IEEE transactions on medical imaging*, vol. 36, no. 2, pp. 674-683, 2016.
- [146] A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei, "What's the point: Semantic segmentation with point supervision," in *European conference on computer vision*, 2016, pp. 549-565: Springer.
- [147] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 618-626.
- [148] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921-2929.
- [149] J. Pont-Tuset, P. Arbelaez, J. T. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping for image segmentation and object proposal generation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 1, pp. 128-140, 2016.
- [150] X. Zhu and A. B. Goldberg, "Introduction to semi-supervised learning," *Synthesis lectures on artificial intelligence and machine learning*, vol. 3, no. 1, pp. 1-130, 2009.
- [151] H. Kervadec, J. Dolz, M. Tang, E. Granger, Y. Boykov, and I. B. Ayed, "Constrained-CNN losses for weakly supervised segmentation," *Medical image analysis*, vol. 54, pp. 88-99, 2019.
- [152] A. B. Watson, "A formula for human retinal ganglion cell receptive field density as a function of visual field location," *Journal of vision*, vol. 14, no. 7, pp. 15-15, 2014.

- [153] E. W. Dijkstra, "A note on two problems in connexion with graphs," *Numerische mathematik*, vol. 1, no. 1, pp. 269-271, 1959.
- [154] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Iccv*, 1998, vol. 98, no. 1, p. 2.
- [155] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.
- [156] A. G. Bennett, A. R. Rudnicka, and D. F. Edgar, "Improvements on Littmann's method of determining the size of retinal features by fundus photography," *Graefe's archive for clinical and experimental ophthalmology*, vol. 232, no. 6, pp. 361-367, 1994.
- [157] C. A. Curcio and K. A. Allen, "Topography of ganglion cells in human retina," *Journal of comparative Neurology*, vol. 300, no. 1, pp. 5-25, 1990.
- [158] J. C. Blanks, Y. Torigoe, D. R. Hinton, and R. H. Blanks, "Retinal pathology in Alzheimer's disease. I. Ganglion cell loss in foveal/parafoveal retina," *Neurobiology of aging*, vol. 17, no. 3, pp. 377-384, 1996.
- [159] M. Pavlidis, T. Stupp, M. Hummeke, and S. Thanos, "Morphometric examination of human and monkey retinal ganglion cells within the papillomacular area," *Retina*, vol. 26, no. 4, pp. 445-453, 2006.
- [160] R. W. Rodieck, K. Binmoeller, and J. Dineen, "Parasol and midget ganglion cells of the human retina," *Journal of Comparative Neurology*, vol. 233, no. 1, pp. 115-132, 1985.
- [161] J. Stone and E. Johnston, "The topography of primate retina: A study of the human, bushbaby, and new-and old-world monkeys," *Journal of Comparative Neurology*, vol. 196, no. 2, pp. 205-223, 1981.
- [162] D. M. Dacey, "The mosaic of midget ganglion cells in the human retina," *Journal of Neuroscience*, vol. 13, no. 12, pp. 5334-5355, 1993.

- [163] M. Watanabe and R. Rodieck, "Parasol and midget ganglion cells of the primate retina," *Journal of Comparative Neurology*, vol. 289, no. 3, pp. 434-454, 1989.
- [164] Z. Liu, D. Hammer, and O. Saeedi, "Multimodal adaptive optics imaging of ganglion cells in patients with primary open angle glaucoma," *Investigative Ophthalmology & Visual Science*, vol. 60, no. 9, pp. 4608-4608, 2019.
- [165] M. A. Abozaid, C. S. Langlo, A. M. Dubis, M. Michaelides, S. Tarima, and J. Carroll, "Reliability and repeatability of cone density measurements in patients with congenital achromatopsia," in *Retinal Degenerative Diseases*: Springer, 2016, pp. 277-283.
- [166] D. Cunefare, A. L. Huckenpahler, E. J. Patterson, A. Dubra, J. Carroll, and S. Farsiu, "RAC-CNN: multimodal deep learning based automatic detection and classification of rod and cone photoreceptors in adaptive optics scanning light ophthalmoscope images," *Biomedical optics express*, vol. 10, no. 8, pp. 3815-3832, 2019.
- [167] G. Wang *et al.*, "Interactive medical image segmentation using deep learning with image-specific fine tuning," *IEEE transactions on medical imaging*, vol. 37, no. 7, pp. 1562-1573, 2018.
- [168] S. Majumder and A. Yao, "Content-aware multi-level guidance for interactive instance segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11602-11611.