eNeuro

Research Article: Methods/New Tools | Novel Tools and Methods

A semi-supervised pipeline for accurate neuron segmentation with fewer ground truth labels

https://doi.org/10.1523/ENEURO.0352-23.2024

Received: 9 September 2023
Revised: 21 December 2023
Accepted: 4 January 2024

Manuscript Title: A semi-supervised pipeline for accurate neuron segmentation with fewer ground truth labels

Abbreviated Title: Semi-supervised neuron segmentation at few labels

Casey M. Baker[1], Yiyang Gong[1,2]
[1]Department of Biomedical Engineering, Duke University, Durham, North Carolina, United States of America, [2]Department of Neurobiology, Duke University, Durham, North Carolina, United States of America

Author Contributions: CB and YG designed research; CB performed research and analyzed data; CB and YG wrote the manuscript.

Correspondence should be addressed to: casey.baker@duke.edu

Number of Figures: 20

Number of Tables: 9

Number of Multimedia: 0

Number of words for Abstract: 189

Number of words for Significance Statement: 120

Number of words for Introduction: 748

Number of words for Discussion: 1314

Acknowledgements:

Conflict of Interest – Authors report no conflict of interest.

Funding sources – This work was funded by the NIH NINDS (1DP2-NS111505) New Innovator Award (Y.G.), the NIH NINDS (1UF1-NS107678) BRAIN Initiative (Y.G.), and the NSF GFRP (C.M.B.).

Abstract

Recent advancements in two-photon calcium imaging have enabled scientists to record the activity of thousands of neurons with cellular resolution. This scope of data collection is crucial to understanding the next generation of neuroscience questions, but analyzing these large recordings requires automated methods for neuron segmentation. Supervised methods for neuron segmentation achieve state of-the-art-accuracy and speed, but currently require large amounts of manually generated ground truth training labels. We reduced the required number of training labels by designing a semi-supervised pipeline. Our pipeline used neural network ensembling to generate pseudolabels to train a single shallow U-Net. We tested our method on three publicly available datasets and compared our performance to three widely-used segmentation methods. Our method outperformed other methods when trained on a small number of ground truth labels and could achieve state-of-the-art accuracy after training on approximately a quarter of the number of ground truth labels as supervised methods. When trained on many ground truth labels, our pipeline attained higher accuracy than that of state-of-the-art methods. Overall, our work will help researchers accurately process large neural recordings while minimizing the time and effort needed to generate manual labels.

Significance statement

Modern neuroscience analyzes the activity of hundreds to thousands of neurons from large optical imaging datasets. One important step in this analysis is neuron segmentation. Supervised algorithms have performed neuron segmentation with class-leading accuracy and speed but lag unsupervised algorithms in training time. A large component of training time is the manual labeling of neurons as training samples; current supervised methods train over many manual labels to achieve accurate prediction. We developed a semi-supervised neuron segmentation algorithm, SAND, that retained high accuracy in the few-label regime. SAND employed neural network ensembling to generate robust pseudolabels and used domain-specific hyperparameter optimization. SAND was more accurate than existing supervised and unsupervised algorithms in low and high label regimes of multiple imaging conditions.

Introduction

Studying modern neuroscience questions requires scientists to simultaneously measure and analyze the coordinated activity of neural ensembles formed from hundreds to thousands of neurons (Makino et al., 2017; Rumyantsev et al., 2020; Stevenson & Kording, 2011; Stringer et al., 2019; Vyas et al., 2020; Yuste, 2015). Understanding the function of neural ensembles is technically challenging because distinctive genetic or functional sub-types of neurons within ensembles spatially overlap and temporally change on timescales ranging from seconds to days (Driscoll et al., 2017; Pérez-Ortega et al., 2021; Sweis et al., 2021; Ziv et al., 2013) .

Calcium imaging using fluorescent protein sensors meets these technical recording challenges because it can record neural ensembles with cellular spatial resolution and genetic specificity over multiple months (Chen et al., 2013; Nakai et al., 2001; Stosiek et al., 2003; Y. Zhang et al., 2023). Calcium influx follows action potentials and typically increases the brightness of calcium indicators (Grienberger & Konnerth, 2012). Recent optical setups have successfully

recorded the calcium activity of hundreds of thousands of neurons simultaneously (Demas et al., 2021). Modern calcium protein sensors have trended toward detection of single action potentials and linear response over multiple action potentials (Ryan et al., 2023; Y. Zhang et al., 2023).

Cellular or sub-cellular resolution imaging that captures rapid single-spike calcium transients creates large datasets. Extracting single neuron activity from these large-scale imaging datasets necessitates a pipeline of automated methods; such algorithms could save time and minimize human error during analysis (Stevenson & Kording, 2011). Analysis pipelines usually consist of four steps to predict spiking activity from calcium fluorescence recordings: 1) motion correction, 2) cell segmentation, 3) fluorescence extraction, and 4) spike inference (Bao et al., 2022; Giovannucci et al., 2019; Keemink et al., 2018; Pachitariu et al., 2017; Pnevmatikakis & Giovannucci, 2017; Theis et al., 2016). Automated neuron segmentation in particular has received substantial attention, but needs improvement.

Both supervised and unsupervised machine learning methods exist for neuron segmentation (Abbas & Masip, 2022; Bao & Gong, 2023). Supervised methods consist of convolutional neural networks (CNNs), while unsupervised methods include dictionary learning, PCA/ICA, and matrix factorization (e.g. CaImAn (Giovannucci et al., 2019) and Suite2p (Pachitariu et al., 2017)). Supervised methods are typically more accurate than unsupervised methods (Bao et al., 2021; Soltanian-Zadeh et al., 2019). For example, Shallow U-Net Neuron Segmentation (SUNS) is a supervised deep learning-based pipeline for neuron segmentation that achieves state-of-the-art accuracy and speed (Bao et al., 2021).

Supervised methods trade off superior performance for the large effort required to generate hundreds of ground truth labels for model training and hyperparameter optimization. The many manual labels can help train algorithms that account for idiosyncratic fluorescence and noise distributions within each image dataset, but then necessitate labels for each imaging condition. Generating such labels is time consuming and subject to human error (Giovannucci et al., 2019; Zhang et al., 2020).

Semi-supervised learning presents an opportunity to reduce the burdens of manual labeling. Semi-supervised segmentation leverages limited numbers of ground truth labels and unlabeled images to train models using two primary approaches: pseudolabeling and consistency regularization (Ouali et al., 2020). Pseudolabeling increases the size of the training dataset by accepting high-confidence labels predicted on unlabeled data as ground truth labels that can further train the model (Lee, 2013; Zou et al., 2020). Consistency regularization trains models by penalizing dissimilar predictions for similar inputs (Chaitanya et al., 2020; Huang et al., 2022; Wu et al., 2022; Zhuang et al., 2021). A combination of pseudolabeling and consistency regularization significantly improved classification accuracy with small numbers of ground truth labels (Sohn et al., 2020).

An alternative paradigm to semi-supervised learning that improves generalizability is ensemble learning. Ensemble learning improves predictive accuracy by combining the outputs of multiple models (Sagi & Rokach, 2018). Averaging multiple independent models reduces overfitting, increases generalizability, and compensates for high model variability even when trained on limited data (Dietterich, 2002; Polikar, 2006). Previous work has successfully applied ensemble learning to neural networks for image classification and segmentation (Muller et al., 2022; Zheng et al., 2019), with the ensemble outperforming the individual (Krizhevsky et al., 2017).

In this study, we developed a semi-supervised neuron segmentation pipeline that maintained state-of-the-art accuracy and prediction speed while limiting the number of manual training labels. Our approach, Semi-supervised Active Neuron Detection (SAND), used neural network ensemble learning to predict active neurons in unlabeled frames. These predictions acted as pseudolabels to augment our training set. We also developed a novel pipeline to choose algorithm hyperparameters with few ground truth labels.

Materials and Methods

Our SAND approach consisted of three main steps: 1) pre-processing the entire video to enhance active neurons, 2) semi-supervised CNN training using small numbers of manually-labeled frames, and 3) post-processing to segment unique neuron masks from the CNN output (**Figure 1A**). The post-processing step used four hyperparameters. Their values were determined using only the manually-labeled frames.

Pre-processing

Before training, we pre-processed the video to reduce noise and emphasize active neurons. We first applied pixel-by-pixel temporal filtering to the registered video, which highlighted fluorescence activity that was similar to calcium response waveforms (Bao et al., 2021). We convolved each pixel with the time-reversed average fluorescence response of the ground truth neurons. Selected fluorescence responses had a peak SNR between 5 and 8, and we aligned the transients by their peaks. We then diminished nonresponsive neurons and enhanced active neurons by converting the temporally-filtered video into an SNR representation. We calculated this representation by first computing the pixel-wise median image and quantile-based noise image over the entire video. We then pixel-wise subtracted the median image from each frame and pixel-wise divided the result by the noise image.

Model training

The original SUNS training pipeline used a fully-supervised approach and trained a single shallow U-Net with a combination of dice loss and focal loss (Bao et al., 2021). The CNN predicted probability maps that underwent a post-processing pipeline to calculate the final neuron masks. Our SAND approach used neural network ensembling to generate pseudolabels (**Figure 1B**). We used an ensemble of three models based on recent work that developed a semi-supervised pipeline for accurate medical image segmentation that trained an ensemble of the same size (Wu et al., 2022). We first defined three separate shallow U-Nets. Each U-Net had a unique decoder

architecture, and one U-Net had the same architecture as SUNS (**Figure 1-1**). We selected the three U-Net architectures tested by Bao et al. (2021) that achieved the highest accuracy. We trained all three U-Nets on frames with manually labeled masks using a weighted sum of dice and focal loss for 200 epochs (Focal loss:Dice loss = 100:1) (**Figure 1-2A**). We then passed 1800 unlabeled frames through each trained U-Net within the ensemble and averaged the output probability maps to serve as pseudolabels. Pseudolabels closely resembled the known temporal masks (**Figure 1-3A-B**). We then produced the final prediction U-Net by using the pseudolabels to continue training the U-Net with the SUNS architecture. We trained this U-Net using binary cross entropy loss for 25 epochs using the pseudolabels (**Figure 1-2A**), and then we fine-tuned the U-Net with a final round of training using dice and focal loss for 200 epochs using the original labeled frames (**Figure 1-2A**). Training time increased as the number of labeled training frames increased but remained under an hour for up to 500 training frames (**Figure 1-2B**). For all training steps, we used the Adam optimizer with a 0.001 learning rate, and our training pipeline augmented the input frames with random flips and rotations to help prevent overfitting.

Post-processing

The output probability maps of our neural network represented the model's confidence that a pixel belonged to an active neuron. Additional post-processing converted the output series of probability maps into unique neuron masks (**Figure 1C**). We followed the same post processing steps described in Bao et al. (Bao et al., 2021). First, we binarized the probability maps with a probability threshold (*p_thresh*) to determine active pixels. Higher values of *p_thresh* retained only high-confidence predictions. Lower values preserved lower confidence predictions, such as pixels from neurons with relatively low SNR, but also kept more false positive predictions. After probability thresholding, we grouped active pixels within a frame into separate components using connected component labeling. We removed components smaller than a minimum area (*min_area*), as these regions were unlikely to be neurons. Next, we merged co-localized

components across different frames; active components in the same location across multiple frames likely represented the same neuron. We defined components as colocalized if the centers of mass (COMs) of two components were within a minimum threshold (COM distance < *centroid_dist*) or if the areas of two components were substantially overlapping. Overlapped neurons met either of two criteria: 1) intersection-over-union (IoU) > 0.5, or 2) consume ratio (consume) > 0.75), with IoU and consume defined for two binary masks m_1 and m_2 as follows (Bao et al., 2021):

$$\mathrm{IoU} = \frac{|m_1 \cap m_2|}{|m_1 \cup m_2|},$$

$$\mathrm{consume} = \frac{|m_1 \cap m_2|}{m_2}.$$

These temporally merged components represented unique ROIs. Lastly, we removed masks that were not active for a minimum number of consecutive frames (*min_consecutive*) typical of calcium responses.

Hyperparameter optimization

Selection of the optimal postprocessing hyperparameters after CNN training was crucial for accurately identifying neurons and distinguishing neurons from noise. Hyperparameter optimization with SUNS required manual labeling of all active neurons in the training video. The original SUNS pipeline used a grid search to determine the postprocessing parameters that maximized F_1 on the training frames (**Table 1-1**). Recall, precision, and F_1, are common metrics to define segmentation accuracy:

$$\mathrm{Recall} = \frac{\#\ \mathrm{True\ Positives}}{\#\ \mathrm{Ground\ Truth}}$$

$$\mathrm{Precision} = \frac{\#\ \mathrm{True\ Positives}}{\#\ \mathrm{Predicted}}$$

$$F_1 = \frac{2}{\mathrm{Recall}^{-1} + \mathrm{Precision}^{-1}}$$

198 Evaluating F_1 on an entire video is impossible using a video that contains unlabeled neurons,
199 which could be inaccurately labeled as false positives. Similarly, the nature of the
200 *min_consecutive* hyperparameter required all video frames to be used in its estimation. We found
201 that a grid search failed to find the optimal hyperparameters when trained with a small number of
202 labels. In particular, we found that a grid search often underestimated the optimal *p_thresh* value
203 when trained with limited manually labeled frames (**Figure 1-4A**).

204 We developed a novel pipeline, Few Label Hyperparameter Optimization (FLHO), to
205 optimize postprocessing hyperparameters that used only a fraction of the number of ground truth
206 labels as SUNS (**Figure 1-5**). Instead of using a grid search to determine all four
207 hyperparameters, we directly calculated *p_thresh* and *min_consecutive* using estimates from a
208 small number of ground truth labels.

209 We first used the ground truth labels to estimate *p_thresh* (**Figure 1-5A**). For each labeled
210 neuron, we identified the frames when that neuron's peak SNR (pSNR) exceeded the threshold
211 set by Bao et al. (2021). The trained CNN then calculated probability maps for these active frames.
212 For each neuron, we found the median probability map value within its mask during its active
213 frames. We used this distribution of median probability values for each neuron to find two values:
214 1) the 25th percentile, which was used for intermediate steps, and 2) the median, which was used
215 as the final *p_thresh*. We used a lower *p_thresh* for intermediate steps that used only labeled
216 frames because our initial small set of labeled training frames likely did not include the frames
217 with the pSNR or peak probability values for each neuron. Our 25th percentile value for *p_thresh*
218 thresholded probability maps and retained neurons with relatively low SNR on the training frames
219 **(Figure 1-3C)**. We used these thresholded maps to perform a grid search for values of
220 *centroid_dist* and *min_area* that maximized the F_1 score on the labeled frames (**Figure 1-5B**).

221 We found that the pipeline was robust across different choice of percentiles with respect
222 to the ultimate algorithm accuracy (**Figure 1-4B**). The values of *centroid_dist* and *min_area* were
223 also robust to changes in *p_thresh*, which may partially explain the robustness in accuracy across

percentiles (**Figure 1-4C**). Additionally, the median value of the *p_thresh* distribution trained on a small number of labels was very similar to the optimal *p_thresh* value calculated using all labels (**Figure 1-4A**). Therefore, we set our final *p_thresh* to the median value. We set an upper bound on this value so that *p_thresh* was not greater than 80% probability. Finally, we calculated *min_consecutive* by assessing the distribution of consecutive frames for all neurons (**Figure 1-5C**). For this step, we used the probability maps for all frames. Therefore, we set *p_thresh* to its final (median) value. We thresholded these probability maps using *p_thresh* and *min_area*. We calculated the maximum number of consecutive frames that the model identified for each neuron. We observed that the minimum consecutive frame value among all neurons was occasionally an outlier, so we selected the second smallest value to be *min_consecutive*. However, the performance of our method was robust across different choices of *min_consecutive* (**Figure 1-4D**). We set an upper bound on *min_consecutive* so that it did not surpass 8 frames (**Figure 1-5D**).

Peer segmentation methods

SUNS: Shallow U-Net Neuron Segmentation (SUNS) is a supervised deep learning pipeline for neuron segmentation from fluorescence recordings (Bao et al., 2021). SUNS first computed an SNR representation of imaging videos that emphasized active neurons and de-emphasized inactive neurons. SUNS then trained a shallow U-Net on 1800 to 2400 of all imaging frames developed from a set of comprehensively labeled neurons over all imaging movies. Finally, a multi-step post-processing pipeline identified unique ROIs across all frames. SUNS determined the hyperparameters for this post-processing pipeline with a grid search that evaluated accuracy against the ground truth labels. Python code for SUNS is available at https://github.com/YijunBao/Shallow-UNet-Neuron-Segmentation_SUNS.

CaImAn: CaImAn is a calcium imaging analysis pipeline that uses both unsupervised and supervised algorithms to identify active neurons (Giovannucci et al., 2019; Pnevmatikakis et al., 2016). The unsupervised step was a non-negative matrix factorization method that separated spatially overlapping neurons based on the temporal activity of active neurons; these sparse decomposed components also included sources that represented background noise and neuropil activity. Components representing unique regions of interest (ROIs) were curated by iteratively combining components that exceeded a threshold for correlated temporal activity. The supervised portion was a quality control step to remove nonneuronal components. This step used a peak signal-to-noise (SNR) threshold, spatial footprint consistency, and a CNN classifier. Python code for CaImAn is available at https://github.com/flatironinstitute/CaImAn (version 1.6.4).

Suite2p: Suite2p is another widely used pipeline that applies unsupervised algorithms to identify potential neurons and a supervised quality control step to refine the neurons (Pachitariu et al., 2017). Suite2p first reduced the dimensionality of the input video using singular value decomposition. Then, unsupervised nonnegative matrix factorization identified ROIs and modeled decomposed neural activity as the weighted sum of underlying neural activity and neuropil signal. A supervised classifier then processed these ROIs and separated cells from non-cells based on temporal and spatial features. Lastly, manual acceptance or rejection of the classifier's predictions refined the final output neurons. Python code for Suite2p is available at https://github.com/MouseLand/suite2p (version 0.6.16).

Datasets

We tested our pipeline on two-photon videos from three different datasets, all recorded in mice. These videos covered multiple cortical and subcortical brain regions, were collected with multiple imaging conditions, and utilized various calcium sensors with different responses and kinetics (**Table 1-2**).

Allen Brain Observatory: The dataset from the Allen Brain Observatory (ABO) consisted of 10 videos recorded from a depth of 275 μm and 10 videos recorded from a depth of 175 μm in the primary visual cortex (V1) (de Vries et al., 2020). The 175 μm set had ~200 neurons per video, and the 275 μm set had ~300 neurons per video. For each depth, we used 10-fold cross validation: we trained our model and determined the hyperparameters using one video and tested on the other nine videos. Data is available at http://observatory.brain-map.org/visualcoding.

Neurofinder: We used three sets of videos (01, 02, and 04) from three different labs with different imaging conditions from the Neurofinder competition (CodeNeuro, 2016). Each video was paired with another video obtained under the same imaging conditions, making 6 pairs of videos. For each of the 6 pairs, we trained the model and determined the hyperparameters on one video and tested on the other video. The 12 videos averaged ~250 neurons per video. Videos are available at http://Neurofinder.Codeneuro.Org/.

CaImAn: The CaImAn dataset (Giovannucci et al., 2019) contained four videos (J115, J123, K53, and YST) that imaged various brain regions. We divided each video into quarters to perform cross validation, so that the training and test set had the same imaging conditions. For two of the videos (J115 and K53), the average number of neurons per sub-video was ~200. For these videos, we trained the model on one sub-video and tested on the remaining three sub-videos. The other two videos (J123 and YST) had ~40 and ~80 neurons per sub video, respectively. For these videos containing far fewer neurons, we used leave-one-out cross validation, training on three sub-videos and testing on the remaining sub-video.

Analysis

We compared three different deep learning segmentation pipelines: 1) SUNS: model training with supervised learning (SL) and hyperparameter optimization with a full grid search (GS), 2) SL and our new hyperparameter optimization pipeline (FLHO), and 3) SAND: model training using a combination of SL and neural ensemble learning, and hyperparameter optimization with FLHO. We also compared SAND to the widely used matrix factorization methods

Suite2p and CaImAn. We quantified the quality of the identified masks as the proportion of the mask's area divided by the area of the mask's convex hull. We evaluated model accuracy by calculating the F_1 score of each method on the test videos when trained with different numbers of ground truth neuron masks from the training video. We altered the number of ground truth masks used in training by randomly sampling different sets of SNR frames (**Figure 1-6**). We evaluated F_1 across all frames and neurons in the test videos using the same ground truth masks as previous work (Bao et al., 2021; Soltanian-Zadeh et al., 2019). For CaImAn and Suite2p, we used the F_1 values found in (Bao et al., 2021), which previously optimized the hyperparameters for these pipelines.

We ran multiple analyses to test the performance of SAND. First, we compared SAND to SUNS, SL + FLHO, CaImAn, and Suite2p when trained on a low number of ground truth neurons. We also compared the performance of SAND trained on a low number of ground truth neurons to the asymptotic performance of SUNS. Finally, we compared the asymptotic performance of SAND to the asymptotic performance of SUNS. We binned the F_1 scores for each condition by the number of neurons used in training. We compared algorithms using the Wilcoxon rank-sum test and by computing the effect size (Cohen's *d*).

Code Accessibility

The code described in the paper is available at https://github.com/caseymbaker/SemiSupervisedNeuronSegmentation2p. The github repository includes a tutorial for downloading and running SAND as well as the data and code for recreating figures in this paper. Code was tested on two Windows 10 PCs (AMD Ryzen 9 3900X, 128 GB RAM, RTX 2080 and Intel Core i7-7700K, 64 GB RAM, Quadro P5000).

Results

We first evaluated SAND using both ABO datasets (**Figure 2**). Masks generated by SAND closely matched the ground truth masks even when trained on only 10 frames (**Figure 2A-B**). Masks generated by SUNS trained on few frames, however, included many false positives, and masks generated by Suite2p and CaImAn were more irregularly shaped and less accurate than those generated by SAND (**Figures 2A-B, 2-1A; Table 2-1, 2-2**). SAND significantly outperformed all other methods when trained on 0-50 ground truth labels (~10 labeled frames) (**Figure 2-2**; **Table 2-1**). In the 275 μm dataset, SUNS achieved a median F_1 score of 0.81 when trained on more than 250 labels (**Figure 2C; Table 2-1**). However, SAND achieved this F_1 score when trained on only ~25% the number of labels and came within one standard deviation of this value when trained on only ~12% the number of labels (median F_1 = 0.79, 34 ± 10 neurons). Additionally, the F_1 score for SAND when trained on more than 250 neurons was significantly higher than the SUNS F_1 score (**Table 2-1**). In the 175 μm dataset (**Figure 2D**), SUNS achieved a median F_1 score of 0.81 when trained on more than 200 neuron labels (**Table 2-1**). However, SAND came within one standard deviation of this value when trained on only ~13% the number of labels (median F_1 = 0.77, 29 ± 12 neurons). Additionally, the F_1 score for SAND when trained on more than 200 labels was significantly higher than the SUNS F_1 score when trained on more than 200 labels (**Table 2-1**). SAND also significantly outperformed the matrix factorization methods, CaImAn and Suite2p, over all numbers of ground truth masks (**Table 2-1**). In particular, SAND trained on only 10 frames more reliably detected low pSNR neurons than CaImAn and Suite2p (**Figure 2-3**). SAND generally improved model precision (**Figure 2C-D**). Both our new training method and our new hyperparameter optimization method helped maximize F_1 in our pipeline. FLHO without pseudolabel training (SL + FLHO) had a modest effect on accuracy when trained on fewer ground truth masks (**Figure 2C-D**). In addition to state-of-the-art accuracy, SAND

also achieved the state-of-the-art processing speed of SUNS at ~300 frames per second **(Figure 2-4)**.

We next tested SAND on the Neurofinder dataset (**Figure 3**). Masks generated by SAND closely matched the ground truth masks even when trained on only 10 frames (**Figure 3A-B**). Masks generated by Suite2p and CaImAn were more irregularly shaped and had more false negative predictions than SAND (**Figures 3A-B, Figure 2-1B, Table 2-3**). SAND significantly outperformed SUNS when trained on 0-50 ground truth neuron labels (~10 frames) (**Figure 3C, Figure 3-1; Table 3-1**). SUNS achieved a median F_1 score of 0.58 when trained on 200-250 labels. However, the performance of SAND was not significantly different from this when trained on only ~14% the number of labels (median F_1 = 0.53, 32 ± 12 neurons; **Table 3-1**). Similar to observations when processing the ABO datasets, our new hyperparameter optimization without pseudolabel training partially improved accuracy when trained on fewer ground truth masks. SAND performed as well as or better than CaImAn segmentation over all numbers of ground truth masks (**Table 3-1**). Overall, the Neurofinder dataset had the most variability in performance, likely due to the variety of imaging conditions throughout this dataset.

Finally, we tested SAND on the CaImAn dataset, starting with the K53 and J115 videos (**Figure 4A-B**). When processing the K53 dataset, SAND significantly outperformed SUNS, Suite2p, and CaImAn at all numbers of ground truth neurons (**Table 4-1**). SAND's performance when trained on 0-50 neurons (~10-25 frames) was more accurate than the performance of SUNS using more than 150 ground truth neurons (~500-1800 frames) **(Figure 4A, Figure 4-1; Table 4-1).** When processing the J115 dataset, SAND significantly outperformed CaImAn and Suite2p on all numbers of ground truth neurons (**Figure 4B**, **Table 4-1**). SAND also significantly outperformed SUNS when trained on 0-50 ground truth neuron labels (~10 frames) **(Figure 4-1; Table 4-1)**. For both videos, SAND's predicted masks aligned closely with the ground truth masks, even when trained on just 10 frames (**Figure 4-2**). SUNS's predicted masks included many false positives. Conversely, CaImAn and Suite2p both failed to detect many ground truth neurons. SAND

outperformed CaImAn and Suite2p on both the YST and J123 videos on all numbers of ground truth neurons; however, SAND did not consistently outperform SUNS (**Figure 4C-D**, **Figure 4-3**). On all of the CaImAn videos, SAND predicted masks with more consistent soma shapes than other methods (**Figure 2-1C, Table 2-4**).

To understand why SAND only moderately outperformed SUNS when processing the J123 and YST videos, we compared the quality of these videos to the quality of the other datasets. The pSNR of a neuron's fluorescence can predict likeliness of being detected by both supervised and unsupervised segmentation methods: neurons with higher pSNR were more likely to be detected (Bao et al., 2021). We calculated the average and standard error of pSNR for all ground truth neurons in each video (**Figure 4-4**).

Neurons in J123 and YST had both lower average pSNR and more variable pSNR than neurons in other videos. This suggests that SAND works best on videos with high pSNR values and low variability of pSNR across neurons. However, SAND appears to be effective when only one of these conditions is met. For example, SAND effectively processed video K53, which had high pSNR but high variability; it also effectively processed the Neurofinder dataset, which had low variability but low pSNR.

The type of calcium indicator used in each recording impacted the pSNR values. Notably, the J123 and YST videos used GCaMP5 (Akerboom et al., 2012) and GCaMP3 (Tian et al., 2009), respectively. These older sensors have very low SNR relative to modern sensors, such as the GCaMP6 used in the other videos (**Table 4-2**). Protein sensors of calcium have continued to develop, so recent sensors in the GCaMP8 series have even higher SNR than that of GCaMP6 (Chen et al., 2013; Ryan et al., 2023; Y. Zhang et al., 2023). It is likely that the high SNR of modern sensors will translate to high pSNR in two-photon neural recordings. This superior signal fidelity should more effectively allow our pipeline to accurately process modern neural recordings with small numbers of ground truth labels.

Finally, we tested how different imaging conditions (e.g. pSNR variability) affected the generalizability of SAND (**Figure 4-5**). We found that SAND generalized well when the training and test data had similar imaging conditions. For example, SAND trained on the ABO 175 µm dataset and tested on the ABO 275 µm dataset performed as well as SAND trained on the ABO 275 µm dataset and tested on the ABO 275 µm dataset. We then tested ABO-trained SAND on the K53 dataset, which has higher average pSNR values and higher pSNR variability than the ABO dataset. We found that ABO-trained SAND still outperformed CaImAn and Suite2p on K53, but K53-trained SAND achieved the highest accuracy across all numbers of training labels. The accuracy of SAND and SUNS trained on the ABO 275 µm dataset and tested on the K53 dataset decreased as the number of ABO labels used to train these models increased. This is likely the result of increased model specificity when trained on data specific to certain imaging conditions. Augmenting the training data of SAND to make consistent predictions on a variety of noise levels would likely improve model generalizability. For example, we could add an additional training step to SAND to include mutual consistency learning: we could train SAND to predict the same probability maps after adding different amounts of noise to the same frame.

Discussion

Current methods of neuron segmentation have a trade-off between accuracy and manual effort: supervised methods have superior accuracy but require substantial manual effort to generate ground truth labels for each imaging condition (Abbas & Masip, 2022). This work developed SAND, the first semi-supervised pipeline to segment active neurons from two-photon calcium recordings with limited ground truth labels. SAND effectively operated in this low label regime by using neural network ensembling and a new hyperparameter optimization pipeline. The former process generated a large and robust set of pseudolabels that trained a deep learning segmentation algorithm, while the latter process determined post-processing hyperparameters from limited numbers of ground truth labels.

SAND achieved higher accuracy than the accuracy of fully-supervised methods at multiple scales of labeling. At the small scale, SAND trained on labels from less than 1% of frames and 25% of all ground truth labels available in a movie was comparably accurate as fully-supervised methods trained on all labels. When trained on all available ground truth labels in our movies (more than 200 neurons), SAND attained higher accuracy than that of current methods. SAND trained on low number of ground truth labels also consistently outperformed matrix factorization methods.

The high accuracy of SAND trained on low numbers of manual labels could allow researchers to circumvent the accuracy-effort tradeoff. SAND attained state-of-the-art accuracy with approximately 25% of the manual labels, but likely even lower fractions of labeling effort. Previous studies on supervised methods required the manual labeling of all hundreds to thousands of neurons in a single video to serve as a comprehensive training set (Bao et al., 2021; Soltanian-Zadeh et al., 2019). We estimate that manual labelers could identify and outline a single neuron per minute, with diminishing speed as they find fewer neurons when scanning through more frames of a movie. Therefore, SAND could greatly reduce the labeling time needed to generate effective labels for training deep learning neuron segmentation algorithms to well under one hour per experimental condition.

Pseudolabel training and FLHO both played a role in SAND's high accuracy when trained on few labels. Pseudolabeling generated a robust training dataset much larger than the manually labeled training set. This larger training set helped train our shallow U-net to distinguish between noise and active neurons, reducing the number of false positive calls. On the other hand, FLHO improved accuracy by improving hyperparameters in post-processing. Selection of hyperparameters can greatly impact algorithm performance, but many other pipelines, such as SUNS, CaImAn and Suite2p, employ supervised postprocessing steps that require large numbers of ground truth labels to accurately tune hyperparameters (Bao et al., 2021; Giovannucci et al.,

2019; Pachitariu et al., 2017). FLHO helped bypass the accuracy-effort tradeoff in hyperparameter optimization through direct calculation of certain parameters using the limited ground truth labels.

The relationship between the number of ground truth labels and accuracy of neuron segmentation displayed three trends. First, in the regime of extremely low numbers of labels, such as 20-50, SAND outperformed its fully-supervised sibling SUNS. Second, both algorithms increased F_1 performance as the number of training labels increased, often reaching performance asymptotes at high numbers of labels ranging from 150 to 250 labels. This large number of labels needed to saturate SAND and SUNS highlights the need for large sets of publicly available manual annotations for a variety of data, such that the field can better understand the conditions that saturate neural network-based segmentations. Third, precision often lagged recall in both SUNS and SAND; the increase in precision largely accounted for the increase in F_1. The reason for this is likely two-fold. First, our ensemble learning method averaged the predictions of three models to generate pseudolabels that were conservative, and thus reduced training on samples near the detection threshold that could increase false positives. Second, FLHO was also likely conservative. It produced hyperparameters, such as *p_thresh* values that were higher than those found by grid search on the few-label dataset, which eliminated weakly confident predictions.

SAND reduces the manual labeling effort compared to fully-supervised algorithms, but inherits the prediction speed of the underlying SUNS shallow U-Net architecture **(Figure 2-4).** This speed was an order of magnitude faster than the rate of data collection (Bao et al., 2021). Fast prediction speed can enable researchers to quickly identify neurons of interest from their recordings in real time and perform targeted perturbation experiments within the same imaging session or during imaging. This capability could help researchers study neural ensemble dynamics in memory and perception that are consistent on the minutes time-scale but change from one day to the next (Deitch et al., 2021; Driscoll et al., 2017; Pérez-Ortega et al., 2021; Rule et al., 2020; Ziv et al., 2013). Our ensemble training and hyperparameter optimization processes also reduces training time compared to SUNS because it trained on only 10 to 25 labeled frames,

far fewer than the 1800 frames used for SUNS. The above benefits at training and test time could also arise from partnerships between existing or future neuron segmentation algorithms and our semi-supervised approaches. Because our ensemble learning and FLHO modify the training approach without dictating the underlying supervised machine learning architecture, these training approaches could retain the accuracy or speed of other algorithms while boosting the other algorithms' performance in the low label regime.

Similar to all machine learning neuron segmentation algorithms, SAND will likely benefit from recent developments in protein engineering and video processing. Our work showed that SAND in particular benefits from higher response and small variance in response. Such distributional changes have been instantiated by recent generations of protein calcium indicators, which are both more responsive and more linear (Dana et al., 2019; Y. Zhang et al., 2023). Additionally, the development of novel unsupervised video denoising pipelines, such as DeepInterp (Lecoq et al., 2021) and DeepCAD-RT (Li et al., 2022), may also improve recall by reducing noise, thereby increasing SNR. Increases in pSNR has correlated with increased recall (Bao et al., 2021; Soltanian-Zadeh et al., 2019). SNR gains will likely increase precision as well by reporting even small calcium fluctuations.

Future work could directly improve our implementation of SAND or create alternative implementations. Direct improvement of SAND could optimize the frame selection or model selection to maximize accuracy. Our current approach randomly selected the frames used for labeling. It is possible that systematic selection of these frames could more effectively represent the range of neuron characteristics (e.g. size and pSNR) with even fewer ground truth labels. Additionally, our current approach defaulted to the SUNS shallow U-net architecture as the final neural network to make neuron predictions. Future iterations of SAND could evaluate the accuracy of all ensemble U-nets when processing the ground truth data and then perform pseudolabel training on the U-net with the lowest error. Finally, improvements to SAND or SUNS could also help detect neurons by improving the post-processing classification step. Such

changes could use dynamic information from a large temporal extent to detect sparsely and weakly active neurons (Soltanian-Zadeh et al., 2019).

Application of SAND beyond the two-photon datasets in this work are potentially numerous. Future SAND applications could help process imaging data from one-photon or volumetric imaging settings, which generally have lower SNR than planar two-photon imaging (Ahrens et al., 2013; Ji et al., 2016; Jung et al., 2004; Waters, 2020). SAND can stand alone to process such data, or pair with segmentation algorithms that target specific optical imaging data types (Yuanlong Zhang et al., 2023). Likewise, future testing could also apply SAND to process the diverse calcium recordings of many cell types, such as inhibitory neurons or glia (Akerboom et al., 2013; Mulholland et al., 2021; Semyanov et al., 2020). SAND's ability to accurately segment neurons in the few labels regime can potentially help individual labs process imaging data from distinctive imaging preparations even if a substantial manually labeled training dataset, generated by a single lab or large community, does not yet exist.

526

527

528

529

References

Abbas, W., & Masip, D. (2022). Computational Methods for Neuron Segmentation in Two-Photon Calcium Imaging Data: A Survey. In *Applied Sciences (Switzerland)* (Vol. 12, Issue 14, p. 6876). Multidisciplinary Digital Publishing Institute. https://doi.org/10.3390/app12146876

Ahrens, M. B., Orger, M. B., Robson, D. N., Li, J. M., & Keller, P. J. (2013). Whole-brain functional imaging at cellular resolution using light-sheet microscopy. *Nature Methods 2013 10:5*, *10*(5), 413–420. https://doi.org/10.1038/nmeth.2434

Akerboom, J., Calderón, N. C., Tian, L., Wabnig, S., Prigge, M., Tolö, J., Gordus, A., Orger, M. B., Severi, K. E., Macklin, J. J., Patel, R., Pulver, S. R., Wardill, T. J., Fischer, E., Schüler, C., Chen, T. W., Sarkisyan, K. S., Marvin, J. S., Bargmann, C. I., … Looger, L. L. (2013). Genetically encoded calcium indicators for multi-color neural activity imaging and combination with optogenetics. *Frontiers in Molecular Neuroscience*, *6*(FEB), 43920. https://doi.org/10.3389/FNMOL.2013.00002/BIBTEX

Akerboom, J., Chen, T. W., Wardill, T. J., Tian, L., Marvin, J. S., Mutlu, S., Calderón, N. C., Esposti, F., Borghuis, B. G., Sun, X. R., Gordus, A., Orger, M. B., Portugues, R., Engert, F., Macklin, J. J., Filosa, A., Aggarwal, A., Kerr, R. A., Takagi, R., … Looger, L. L. (2012). Optimization of a GCaMP calcium indicator for neural activity imaging. *Journal of Neuroscience*, *32*(40), 13819–13840. https://doi.org/10.1523/JNEUROSCI.2601-12.2012

Bao, Y., & Gong, Y. (2023). Machine learning data processing as a bridge between microscopy and the brain. *Intelligent Nanotechnology: Merging Nanoscience and Artificial Intelligence*, 399–420. https://doi.org/10.1016/B978-0-323-85796-3.00014-7

Bao, Y., Redington, E., Agarwal, A., & Gong, Y. (2022). Decontaminate Traces From Fluorescence Calcium Imaging Videos Using Targeted Non-negative Matrix Factorization.

Frontiers in Neuroscience, *15*. https://doi.org/10.3389/FNINS.2021.797421

Bao, Y., Soltanian-Zadeh, S., Farsiu, S., & Gong, Y. (2021). Segmentation of Neurons from Fluorescence Calcium Recordings Beyond Real-time. *Nature Machine Intelligence*, *3*(7), 590–600. https://doi.org/10.1038/s42256-021-00342-x

Chaitanya, K., Erdil, E., Karani, N., & Konukoglu, E. (2020). Contrastive learning of global and local features for medical image segmentation with limited annotations. *Advances in Neural Information Processing Systems*, *2020-Decem*, 12546–12558.

Chen, T. W., Wardill, T. J., Sun, Y., Pulver, S. R., Renninger, S. L., Baohan, A., Schreiter, E. R., Kerr, R. A., Orger, M. B., Jayaraman, V., Looger, L. L., Svoboda, K., & Kim, D. S. (2013). Ultra-sensitive fluorescent proteins for imaging neuronal activity. *Nature*, *499*(7458), 295. https://doi.org/10.1038/NATURE12354

CodeNeuro. (2016). *The neurofinder challenge*.

Dana, H., Sun, Y., Mohar, B., Hulse, B. K., Kerlin, A. M., Hasseman, J. P., Tsegaye, G., Tsang, A., Wong, A., Patel, R., Macklin, J. J., Chen, Y., Konnerth, A., Jayaraman, V., Looger, L. L., Schreiter, E. R., Svoboda, K., & Kim, D. S. (2019). High-performance calcium sensors for imaging activity in neuronal populations and microcompartments. *Nature Methods*, *16*(7), 649–657. https://doi.org/10.1038/s41592-019-0435-6

de Vries, S. E. J., Lecoq, J. A., Buice, M. A., Groblewski, P. A., Ocker, G. K., Oliver, M., Feng, D., Cain, N., Ledochowitsch, P., Millman, D., Roll, K., Garrett, M., Keenan, T., Kuan, L., Mihalas, S., Olsen, S., Thompson, C., Wakeman, W., Waters, J., … Koch, C. (2020). A large-scale standardized physiological survey reveals functional organization of the mouse visual cortex. *Nature Neuroscience*, *23*(1), 138–151. https://doi.org/10.1038/S41593-019-0550-9

Deitch, D., Rubin, A., & Ziv, Y. (2021). Representational drift in the mouse visual cortex. *Current Biology*, *31*(19), 4327-4339.e6. https://doi.org/10.1016/j.cub.2021.07.062

Demas, J., Manley, J., Tejera, F., Kim, H., Traub, F. M., Chen, B., & Vaziri, A. (2021). High-Speed, Cortex-Wide Volumetric Recording of Neuroactivity at Cellular Resolution using Light Beads Microscopy. *BioRxiv*, 2021.02.21.432164. https://doi.org/10.1101/2021.02.21.432164

Dietterich, T. . (2002). Ensemble Learning. In *The Handbook of Brain Theory and Neural Networks*. MIT Press.

Driscoll, L. N., Pettit, N. L., Minderer, M., Chettih, S. N., & Harvey, C. D. (2017). Dynamic Reorganization of Neuronal Activity Patterns in Parietal Cortex. *Cell*, *170*(5), 986-999.e16. https://doi.org/10.1016/J.CELL.2017.07.021

Giovannucci, A., Friedrich, J., Gunn, P., Kalfon, J., Brown, B. L., Koay, S. A., Taxidis, J., Najafi, F., Gauthier, J. L., Zhou, P., Khakh, B. S., Tank, D. W., Chklovskii, D. B., & Pnevmatikakis, E. A. (2019). Caiman an open source tool for scalable calcium imaging data analysis. *ELife*, *8*. https://doi.org/10.7554/eLife.38173

Grienberger, C., & Konnerth, A. (2012). Imaging Calcium in Neurons. *Neuron*, *73*(5), 862–885. https://doi.org/10.1016/J.NEURON.2012.02.011

Huang, W., Chen, C., Xiong, Z., Zhang, Y., Chen, X., Sun, X., & Wu, F. (2022). Semi-Supervised Neuron Segmentation via Reinforced Consistency Learning. *IEEE Transactions on Medical Imaging*, *41*(11), 3016–3028. https://doi.org/10.1109/TMI.2022.3176050

Ji, N., Freeman, J., & Smith, S. L. (2016). Technologies for imaging neural activity in large volumes. *Nature Neuroscience 2016 19:9*, *19*(9), 1154–1164.

https://doi.org/10.1038/nn.4358

Jung, J. C., Mehta, A. D., Aksay, E., Stepnoski, R., & Schnitzer, M. J. (2004). In Vivo Mammalian Brain Imaging Using One- and Two-Photon Fluorescence Microendoscopy. *Journal of Neurophysiology*, *92*(5), 3121. https://doi.org/10.1152/JN.00234.2004

Keemink, S. W., Lowe, S. C., Pakan, J. M. P., Dylda, E., Van Rossum, M. C. W., & Rochefort, N. L. (2018). FISSA: A neuropil decontamination toolbox for calcium imaging signals. *Scientific Reports*, *8*(1), 1–12. https://doi.org/10.1038/s41598-018-21640-2

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90. https://doi.org/10.1145/3065386

Lecoq, J., Oliver, M., Siegle, J. H., Orlova, N., Ledochowitsch, P., & Koch, C. (2021). Removing independent noise in systems neuroscience data using DeepInterpolation. *Nature Methods 2021 18:11*, *18*(11), 1401–1408. https://doi.org/10.1038/s41592-021-01285-2

Lee, D.-H. (2013). Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. *Workshop on Challenges in Representation Learning, ICML*, *3*(2), 896.

Li, X., Li, Y., Zhou, Y., Wu, J., Zhao, Z., Fan, J., Deng, F., Wu, Z., Xiao, G., He, J., Zhang, Y., Zhang, G., Hu, X., Chen, X., Zhang, Y., Qiao, H., Xie, H., Li, Y., Wang, H., … Dai, Q. (2022). Real-time denoising enables high-sensitivity fluorescence time-lapse imaging beyond the shot-noise limit. *Nature Biotechnology 2022 41:2*, *41*(2), 282–292. https://doi.org/10.1038/s41587-022-01450-8

Makino, H., Ren, C., Liu, H., Kim, A. N., Kondapaneni, N., Liu, X., Kuzum, D., & Komiyama, T. (2017). Transformation of Cortex-wide Emergent Properties during Motor Learning.

Neuron, *94*(4), 880-890.e8. https://doi.org/10.1016/J.NEURON.2017.04.015

Mulholland, H. N., Hein, B., Kaschube, M., & Smith, G. B. (2021). *Tightly coupled inhibitory and excitatory functional networks in the developing primary visual cortex*. *10*, 72456. https://doi.org/10.7554/eLife

Muller, D., Soto-Rey, I., & Kramer, F. (2022). An Analysis on Ensemble Learning Optimized Medical Image Classification with Deep Convolutional Neural Networks. *IEEE Access*, *10*, 66467–66480. https://doi.org/10.1109/ACCESS.2022.3182399

Nakai, J., Ohkura, M., & Imoto, K. (2001). A high signal-to-noise Ca(2+) probe composed of a single green fluorescent protein. *Nature Biotechnology*, *19*(2), 137–141. https://doi.org/10.1038/84397

Ouali, Y., Hudelot, C., & Tami, M. (2020). An Overview of Deep Semi-Supervised Learning. *ArXiv*.

Pachitariu, M., Stringer, C., Dipoppa, M., Schröder, S., Rossi, L. F., Dalgleish, H., Carandini, M., & Harris, K. D. (2017). Suite2p: beyond 10,000 neurons with standard two-photon microscopy. *BioRxiv*. https://doi.org/10.1101/061507

Pérez-Ortega, J., Alejandre-García, T., & Yuste, R. (2021). Long-term stability of cortical ensembles. *ELife*, *10*. https://doi.org/10.7554/ELIFE.64449

Pnevmatikakis, E. A., & Giovannucci, A. (2017). NoRMCorre: An online algorithm for piecewise rigid motion correction of calcium imaging data. *Journal of Neuroscience Methods*, *291*, 83–94. https://doi.org/10.1016/j.jneumeth.2017.07.031

Pnevmatikakis, E. A., Soudry, D., Gao, Y., Peterka, D. S., Yuste, R., & Correspondence, L. P. (2016). Simultaneous Denoising, Deconvolution, and Demixing of Calcium Imaging Data. *Neuron*, *89*, 285–299. https://doi.org/10.1016/j.neuron.2015.11.037

Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, *6*(3), 21–44. https://doi.org/10.1109/MCAS.2006.1688199

Rule, M. E., Loback, A. R., Raman, D. V., Driscoll, L. N., Harvey, C. D., & O'leary, T. (2020). Stable task information from an unstable neural population. *ELife*, *9*, 1–16. https://doi.org/10.7554/ELIFE.51121

Rumyantsev, O. I., Lecoq, J. A., Hernandez, O., Zhang, Y., Savall, J., Chrapkiewicz, R., Li, J., Zeng, H., Ganguli, S., & Schnitzer, M. J. (2020). Fundamental bounds on the fidelity of sensory cortical coding. *Nature*, *580*(7801), 100–105. https://doi.org/10.1038/S41586-020-2130-2

Ryan, M. B., Churchland, A. K., Gong, Y., & Baker, C. (2023). Fastest-ever calcium sensors broaden the potential of neuronal imaging. *Nature 2023 615:7954*, *615*(7954), 804–805. https://doi.org/10.1038/d41586-023-00704-y

Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. In *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* (Vol. 8, Issue 4). https://doi.org/10.1002/widm.1249

Semyanov, A., Henneberger, C., & Agarwal, A. (2020). Making sense of astrocytic calcium signals — from acquisition to interpretation. In *Nature Reviews Neuroscience* (Vol. 21, Issue 10, pp. 551–564). https://doi.org/10.1038/s41583-020-0361-8

Sohn, K., Berthelot, D., Li, C. L., Zhang, Z., Carlini, N., Cubuk, E. D., Kurakin, A., Zhang, H., & Raffel, C. (2020). FixMatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems*, *2020-Decem*.

Soltanian-Zadeh, S., Sahingur, K., Blau, S., Gong, Y., & Farsiu, S. (2019). Fast and robust active neuron segmentation in two-photon calcium imaging using spatiotemporal deep learning. *Proceedings of the National Academy of Sciences of the United States of*

669 *America*, *116*(17), 8554–8563. https://doi.org/10.1073/pnas.1812995116

670 Stevenson, I. H., & Kording, K. P. (2011). How advances in neural recording affect data
671 analysis. *Nature Neuroscience*, *14*(2), 139–142. https://doi.org/10.1038/nn.2731

672 Stosiek, C., Garaschuk, O., Holthoff, K., & Konnerth, A. (2003). In vivo two-photon calcium
673 imaging of neuronal networks. *Proceedings of the National Academy of Sciences of the*
674 *United States of America*, *100*(12), 7319–7324.
675 https://doi.org/10.1073/PNAS.1232232100/ASSET/B3D9FDB8-4437-42E4-8058-
676 8AADAF92B00B/ASSETS/GRAPHIC/PQ1232232005.JPEG

677 Stringer, C., Pachitariu, M., Steinmetz, N., Carandini, M., & Harris, K. D. (2019). High-
678 dimensional geometry of population responses in visual cortex. *Nature*, *571*(7765), 361–
679 365. https://doi.org/10.1038/S41586-019-1346-5

680 Sweis, B. M., Mau, W., Rabinowitz, S., & Cai, D. J. (2021). Dynamic and heterogeneous neural
681 ensembles contribute to a memory engram. *Current Opinion in Neurobiology*, *67*, 199–206.
682 https://doi.org/10.1016/J.CONB.2020.11.017

683 Theis, L., Berens, P., Froudarakis, E., Reimer, J., Román Rosón, M., Baden, T., Euler, T.,
684 Tolias, A. S., & Bethge, M. (2016). Benchmarking Spike Rate Inference in Population
685 Calcium Imaging. *Neuron*, *90*(3), 471–482.
686 https://doi.org/10.1016/J.NEURON.2016.04.014/ATTACHMENT/262BE9B0-ACBE-4015-
687 8BC3-C21A1BA8AF1D/MMC1.PDF

688 Tian, L., Hires, S. A., Mao, T., Huber, D., Chiappe, M. E., Chalasani, S. H., Petreanu, L.,
689 Akerboom, J., McKinney, S. A., Schreiter, E. R., Bargmann, C. I., Jayaraman, V., Svoboda,
690 K., & Looger, L. L. (2009). Imaging neural activity in worms, flies and mice with improved
691 GCaMP calcium indicators. *Nature Methods*, *6*(12), 875–881.
692 https://doi.org/10.1038/nmeth.1398

693 Vyas, S., Golub, M. D., Sussillo, D., & Shenoy, K. V. (2020). *Computation Through Neural*
694 *Population Dynamics*. https://doi.org/10.1146/annurev-neuro-092619

695 Waters, J. (2020). Sources of widefield fluorescence from the brain. *ELife*, *9*, 1–13.
696 https://doi.org/10.7554/ELIFE.59841

697 Wu, Y., Ge, Z., Zhang, D., Xu, M., Zhang, L., Xia, Y., & Cai, J. (2022). Mutual consistency
698 learning for semi-supervised medical image segmentation. *Medical Image Analysis*, *81*,
699 102530. https://doi.org/10.1016/J.MEDIA.2022.102530

700 Yuste, R. (2015). From the neuron doctrine to neural networks. In *Nature Reviews*
701 *Neuroscience* (Vol. 16, Issue 8, pp. 487–497). Nature Publishing Group.
702 https://doi.org/10.1038/nrn3962

703 Zhang, L., Tanno, R., Xu, M. C., Jin, C., Jacob, J., Ciccarelli, O., Barkhof, F., & Alexander, D. C.
704 (2020). Disentangling human error from the ground truth in segmentation of medical
705 images. *Advances in Neural Information Processing Systems*, *2020-Decem*.

706 Zhang, Y., Rózsa, M., Bushey, D., & , J. Zheng, D. Reep, G. J. Broussard, A. Tsang, G.
707 Tsegaye, R. Patel, S. Narayan, J. X. Lim, R. Zhang, M. B. Ahrens, G. C. Turner, S. S.-H.
708 Wang, K. Svoboda, W. Korff, E. R. Schreiter, J. P. Hasseman, I. Kolb, L. L. L. (2023). Fast
709 and sensitive GCaMP calcium indicators for imaging neural populations. *Nature*, *615*, 884–
710 891.

711 Zhang, Yuanlong, Zhang, G., Han, X., Wu, J., Li, Z., Li, X., Xiao, G., Xie, H., Fang, L., & Dai, Q.
712 (2023). Rapid detection of neurons in widefield calcium imaging datasets after training with
713 synthetic data. *Nature Methods*, *20*(5), 747–754. https://doi.org/10.1038/s41592-023-
714 01838-7

715 Zheng, H., Zhang, Y., Yang, L., Liang, P., Zhao, Z., Wang, C., & Chen, D. Z. (2019). A New

716 Ensemble Learning Framework for 3D Biomedical Image Segmentation. *Proceedings of*
717 *the AAAI Conference on Artificial Intelligence*, *33*(01), 5909–5916.
718 https://doi.org/10.1609/AAAI.V33I01.33015909

719 Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., & Yamins, D. L. K.
720 (2021). Unsupervised neural network models of the ventral visual stream. *Proceedings of*
721 *the National Academy of Sciences of the United States of America*, *118*(3), e2014196118.
722 https://doi.org/10.1073/PNAS.2014196118/SUPPL_FILE/PNAS.2014196118.SAPP.PDF

723 Ziv, Y., Burns, L. D., Cocker, E. D., Hamel, E. O., Ghosh, K. K., Kitch, L. J., Gamal, A. El, &
724 Schnitzer, M. J. (2013). Long-term dynamics of CA1 hippocampal place codes. *Nature*
725 *Neuroscience*, *16*(3), 264–266. https://doi.org/10.1038/nn.3329

726 Zou, Y., Zhang, Z., Zhang, H., Li, C.-L., Bian, X., Huang, J.-B., & Pfister, T. (2020). PseudoSeg:
727 Designing Pseudo Labels for Semantic Segmentation. *ArXiv*.

728
729
730