A Deep Learning Approach for Neuronal Cell Body Segmentation in Neurons Expressing GCaMP using a Swin Transformer

https://doi.org/10.1523/ENEURO.0148-23.2023

Cite as: eNeuro 2023; 10.1523/ENEURO.0148-23.2023
Received: 5 May 2023
Revised: 8 September 2023
Accepted: 9 September 2023

This Early Release article has been peer-reviewed and accepted, but has not been through the composition and copyediting processes. The final version may differ slightly in style or formatting and will contain links to any extended data.

Alerts: Sign up at www.eneuro.org/alerts to receive customized email alerts when the fully formatted version of this article is published.
1. **Manuscript Title:** A Deep Learning Approach for Neuronal Cell Body Segmentation in Neurons Expressing GCaMP using a Swin Transformer

2. **Abbreviated Title:** Swin Transformer Neuronal Cell Body Segmentation

3. **Author Names and Affiliations:**

   Mohammad Shafkat Islam\textsuperscript{a}, Pratyush Suryavanshi\textsuperscript{b}, Samuel M. Baule\textsuperscript{c}, Joseph Glykys\textsuperscript{b,d}, Stephen Baek\textsuperscript{a}

   \textsuperscript{a}School of Data Science, The University of Virginia, Charlottesville, VA 22903, USA;
   
   \textsuperscript{b}Dept. of Pediatrics, Iowa Neuroscience Institute, The University of Iowa, Iowa City, IA 52242, USA;
   
   \textsuperscript{c}Dept. of Biomedical Engineering, The University of Iowa, Iowa City, IA 52242, USA;
   
   \textsuperscript{d}Dept. of Neurology, The University of Iowa, Iowa City, IA 52242, USA;

4. **Author Contributions:** SB and JG designed research and wrote the paper; MSI and PS performed research, analyzed data, and wrote the paper; SMB analyzed the data.

5. **Corresponding authors:** Stephen Baek (baek@virginia.edu) and Joseph Glykys (joseph-glykys@uiowa.edu)

6. **Number of Figures:** 10

7. **Number of Tables:** 4 + Statistical Table

8. **Number of Multimedia:** None

9. **Number of words for Abstract:** 194

10. **Number of words for Significance Statement:** 119

11. **Number of words for Introduction:** 527

12. **Number of words for Discussion:** 978
13. **Acknowledgments:** JG was funded by NIH/NINDS R01NS115800 and the Iowa Neuroscience Institute. SB was funded by NIH/NINDS R01NS115800. This research was partly supported by the computational resources provided by the University of Iowa, Iowa City, Iowa, and by The University of Iowa Hawkeye Intellectual and Developmental Disabilities Research Center (HAWK-IDDRC) P50 HD103556.

14. **Conflict of interest statement:** No, the authors declare no competing financial interests.

15. **Funding Sources:** JG was funded by NIH/NINDS R01NS115800 and the Iowa Neuroscience Institute. SB was funded by NIH/NINDS R01NS115800. This research was partly supported by the computational resources provided by the University of Iowa, Iowa City, Iowa, and The University of Iowa Hawkeye Intellectual and Developmental Disabilities Research Center (HAWK-IDDRC) P50 HD103556.
Abstract

Neuronal cell body analysis is crucial for quantifying changes in neuronal sizes under different physiological and pathological conditions. Neuronal cell body detection and segmentation mainly rely on manual or pseudo-manual annotations. Manual annotation of neuronal boundaries is time-consuming, requires human expertise, and has intra-/inter-observer variances. Also, determining where the neuron’s cell body ends and where the axons and dendrites begin is taxing. We developed a deep-learning-based approach that uses a state-of-the-art shifted windows (Swin) transformer for automated, reproducible, fast, and unbiased 2D detection and segmentation of neuronal somas imaged in mouse acute brain slices by multiphoton microscopy. We tested our Swin algorithm during different experimental conditions of low and high signal fluorescence. Our algorithm achieved a mean Dice score of 0.91, a precision of 0.83, and a recall of 0.86. Compared to two different convolutional neural networks, the Swin transformer outperformed them in detecting the cell boundaries of GCamP6s expressing neurons. Thus, our Swin transform algorithm can assist in the fast and accurate segmentation of fluorescently labeled neuronal cell bodies in thick acute brain slices. Using our flexible algorithm, researchers can better study the fluctuations in neuronal soma size during physiological and pathological conditions.

Significance Statement:

Neuronal cell body partitioning is essential for evaluating the effects of physiological and pathological conditions. Neuronal segmentation is challenging due to the complex morphological structures of neuronal cell bodies and their surroundings. Most current
approaches for detecting and segmenting neurons are based on manual or pseudo-manual annotations of the neuronal boundaries by human experts. These are time-consuming and have intra-/inter-observer variability. Leveraging the current success of vision transformers for general object detection and segmentation tasks, we developed a deep-learning-based approach for automated, fast, robust 2D neuronal cell body segmentation using a state-of-the-art vision transformer (Swin transformer). This approach for neuronal cell body segmentation can assist researchers in evaluating the changes in neuronal cell body sizes under different pathological conditions.
Introduction

The accurate estimation of cellular morphology is essential for identifying cytotoxic edema of neurons and astrocytes in multiple pathological conditions, including brain injury, seizures, hypoxia, and ischemia (Andrew et al., 2007; Glykys et al., 2017; Risher et al., 2009). Despite these critical needs, measuring neuronal cell bodies remains challenging.

Neurons have complex structures with multiple branches, and calculating the changes in somatic and dendritic sizes is arduous. Most common detection methods involve manual or semi-automatic measurement of the neuronal area, which can overlook small but relevant morphological changes. This labor-intensive process also carries a high degree of human bias due to its manual nature. Moreover, it is challenging to determine when a neuronal body ends and when processes like dendrites and axons begin unless unique markers are used. Thus, there is an urgent need for automatic, reproducible, fast, and unbiased approaches to detect and segment neuronal cell bodies with high precision. The lack of such methods impedes understanding pathways that regulate neuronal size during various brain injuries.

Identifying neuronal cell bodies that express diverse biosensors in acute brain slices is taxing. The imaged tissue, commonly between 350 to 450 microns thick, has overlapping neurons across different depth planes, fluorescence intensity variations throughout the thickness of a brain slice, and low contrast between neuronal cell bodies and the neuropil, among others. As a result, current state-of-the-art deep-learning-based cell segmentation techniques (Greenwald et al., 2021; Stringer et al., 2021) cannot be directly applied to segment neuronal boundaries in thick brain slices. A mask region-based convolutional
neural network (Mask R-CNN) (Kirillov et al., 2019) based system was recently developed, called ANMAF, which automatically outlines fluorescent neuronal somatic areas in acute brain slices (Tong et al., 2021). However, biosensors with dim or fluctuating fluorescence or slices with a high neuropil signal can hinder contrast-based detection of neuronal cell bodies, making ANMAF less efficient, probably due to its limited ability to build long-range dependencies and global image contexts.

Vision transformers (Carion et al., 2020; Dosovitskiy et al., 2020) were developed to address CNN’s limited ability to build long-range dependencies and global contexts in images, and they were inspired by transformers used in the Natural Language Processing (NLP) domain (Vaswani et al., 2017). While these vision transformers can construct long-range dependencies and global contexts, their computational complexity is quadratic.

Recently, Swin transformers have achieved state-of-the-art performance in object detection by using a hierarchical feature representation and a shifted window (therefore the acronym Swin) approach while maintaining a linear computational complexity (Lin et al., 2021).

Building upon the success of vision transformers, we developed an automated, fast, accurate, reproducible, and unbiased neuronal cell body segmentation algorithm using 2D two-photon images by optimizing the current state-of-the-art Swin transformer (Lin et al., 2021). Our algorithm detects neurons genetically expressing GCaMP6s, a widely used Ca^{2+}-sensitive-fluorophore (Chen et al., 2013), and stable yellow fluorescent protein (YFP). It also segments neuronal cell bodies in neurons expressing GCaMP6s exposed to excitotoxic insult and with higher detection rates than two different CNN algorithms (Mask-R CNN-based algorithm: ANMAF, and a CNN-based approach: Cellpose). Thus, our Swin
transformer will allow researchers to measure changes in fluorescent neuronal cell bodies during different pathological conditions.

**Materials and Methods**

**Experimental design**

Acute brain slices were prepared from neonatal mice (post-natal day 8-12) expressing neuronal GCaMP6s (C57BL/6J-Tg(Thy1-GCaMP6s) GP4.3Dkim/J Strain #:024275, Jackson Lab), and yellow fluorescent protein (YFP; B6.Cg-Tg(Thy1-YFP)HJrs/J Strain #:003782, Jackson Lab). Mice of both sexes were anesthetized with inhaled isoflurane and decapitated per a protocol approved by The University of Iowa. The brain was removed and placed in ice-cold artificial cerebrospinal fluid (aCSF) containing (in mM) NaCl (120), KCl (3.3), CaCl₂ (1.3), MgCl₂ (2), NaH₂PO₄ (1.25), NaHCO₃ (25), and D-glucose (10) with pH 7.3 - 7.4 when bubbled with carbogen (95% O₂ and 5% CO₂). Coronal brain slices 450 µm thick were cut using a vibratome (Leica VT1000S) while submerged in aCSF containing 2 mM kynurenic acid to block glutamatergic receptors. The brain slices were placed in an interface holding chamber containing aCSF (1.3mM MgCl₂) at room temperature for 30 min, after which the temperature was slowly increased to and maintained at 30 °C. Slices were stored for at least 1-hour before being transferred to the recording chamber. NMDA was obtained from Sigma, and stock solutions were prepared and diluted to 30 µM in aCSF on the experimental day.

**Optical imaging and manual detection**

Two-photon laser scanning microscopy (2PLSM) was used to image neurons expressing genetically encoded fluorophores in layer IV/V of the somatosensory neocortex and CA1
hippocampal region. Acute brain slices were placed in a submerged chamber constantly perfused with aCSF maintained at 30 °C. The location of the sensory neocortex was determined using epifluorescence. 2PLSM imaging was performed using the Bruker Ultima galvo-resonant system using an Olympus BX51WIF upright microscope body with a water immersion objective (20X, 1.0 N.A.). A Ti: sapphire tunable laser (Mai Tai HPDS; Spectra-Physics) generated two-photon excitation (920 nm: GCaMP6s; 860 nm: YFP). Scanning was performed with galvo-mirrors. Emitted light was bandpass filtered at 565 nm using a dichroic mirror (T510lpxrxt, Chroma), and green and yellow emission wavelengths were isolated using specific filters: 525/35 nm (green) and 535/30 nm (yellow). GaAsP or multi-alkali photomultiplier tubes (PMT, Hamamatsu Photonics) were used to acquire the emitted signal. Three-dimensional stacks (3D) of raster scans in the XY plane were imaged at 2 µm intervals with a 512×512-pixel resolution. All images were acquired at 2X digital zoom. All fluorophores were imaged at multiple depths. Ca²⁺ signals in GCaMP6s expressing neurons were imaged during baseline aCSF (low fluorescence), 30 µM NMDA perfusion (10 mins, high fluorescence), and washout (up to 40 mins, mid-high fluorescence; Fig. 1). A similar protocol was used for neurons expressing YFP. Images were background subtracted, smoothened (median filter, radius=2), converted to maximum intensity projections (MIPs, every 10 images), and contrast-enhanced (CLAHE) using ImageJ. Thus, the maximal somatic areas for each neuron over 20 µm depth were represented in a MIP, allowing comparisons of the neuronal maximal areas between different experimental conditions (baseline, NMDA, and washout) (Takezawa et al., 2023). For manual tracing, the contours of the neurons were generated using ImageJ and the Canny edge detector plugin.

Deep neural network training
The MIPs of 20 µm depth were utilized to train and evaluate the performance of the deep neural network. To ensure proper separation of training/validation and testing datasets, the Swin transformer was assessed on the independent testing data, which was not used during training/validation. Thus, we first divided the total dataset into two sets: the training/validation dataset ($D_1$, 75 images) and the testing dataset ($D_2$, 15 images). Due to the labor-intensive manual annotation process (see below), generating exhaustive labels of all the neurons in the dataset is almost impossible. Hence, we generated a synthetic training/validation dataset ($\text{Syn}_{D1}$) from $D_1$ (approximately 700 manually annotated neurons), which was then used to train the neural network architecture, determine the hyperparameters, and optimize the model. Once the hyperparameters were determined and the network architecture was optimized, the approach was evaluated on the independent testing dataset ($D_2$) and an independent synthetic testing dataset ($\text{Syn}_{D2}$), which was never seen by the network before. The Swin transformer was trained on a Linux machine with a single GPU (NVIDIA RTX 5000). However, the approach can also be trained and tested on CPU-only machines. The CNN-based algorithms ANMAF and Cellpose were also trained using the same training/validation dataset ($\text{Syn}_{D1}$), and their performances were evaluated on the same independent testing dataset ($D_2$ and $\text{Syn}_{D2}$).

**Manual tracings of somatic perimeter**

To generate the synthetic dataset for training the deep neural network and to evaluate the performance of the Swin transformer network, two human experts with prior experience manually annotated a subset of prominent somatic regions from the training/validation dataset.
dataset (D1, n=75 Z-stack images, 6 brain slices) and the testing dataset (D2, n=15 Z-stack images, 8 brain slices). Human experts proactively manipulated image intensity and contrast and examined adjacent slices to obtain contextual information to ensure the highest possible performance during manual contouring. These manual tracings were used as “reference standards” (ground truth) to train the deep neural network and quantitatively evaluate the performance of the Swin transformer approach.

Quantitative evaluation of the algorithms

The Dice coefficient was used to measure the pixel-wise overlap between the predicted segmentation and the ground truth, as defined by the following equation:

\[
\text{Dice coefficient} = \frac{2 |X \cap Y|}{|X| + |Y|}
\]

We also used the intersection-over-union (IOU) parameter, also known as the Jaccard index, which is the area of overlap between the predicted segmentation and the ground truth divided by the area of union between the two.

\[
\text{IOU} = \frac{|X \cap Y|}{|X \cup Y|}
\]

The detected neurons were labeled either as true positives (TP) if the instances were correct detections, false positive (FP) if the detection was incorrect (either background, wrong boundaries, or multiple instances being predicted as one neuron), or false negative (FN) if the approach missed the neuron. We also computed precision and recall according to the following equations:

\[
\text{Precision} = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)} + \text{False Positive (FP)}}
\]
Recall = \frac{True Positive (TP)}{True Positive (TP) + False Negative (FN)}

The yield rate was defined by the ratio between the total number of neurons correctly detected by the approach in consideration and the total number of ground truth neurons.

**Statistical analysis**

The normality of distributions was determined using the Shapiro-Wilk, Kolmogorov-Smirnov tests and Q-Q plots. Data are presented as mean ± confidence interval (95% CI). Repeated Measure one-way ANOVA with the Dunnett test was used to compare multiple parametric data to a single group. Two-way ANOVA was used to study the interaction between algorithms and experimental intervention. Statistical significance was considered at p<0.05 (Table 1).

**Code accessibility**

The code described in the manuscript is freely available online at https://github.com/stephenbaek/cell-transformer. The code is available as extended data.

**Results**

**Synthetic data generation**

Training a deep neural network requires an extensive collection of input images with manually annotated labels. However, exhaustive labeling of all instances of somatic boundaries on the neuronal images is highly time-consuming for a human expert. Hence, to train a Swin transformer, we first generated synthetic neurons (Fig. 2). A human expert manually annotated a few GCaMP6s-expressing neurons imaged in acute brain slices during baseline, NMDA perfusion and washout conditions (capturing different fluorescence
intensities and somatic sizes) and non-somatic regions from the input images to generate a collection of background tiles (n=610 cells, 44 backgrounds; Fig. 2A, B). The background tiles were placed randomly and blended to avoid unnatural contrast between neighboring tiles to create a collection of synthetic background images (Fig. 2C). Finally, the manually identified somatic boundaries were pasted into randomly selected synthetic background images to produce a collection of synthetic cell images (Fig. 2D). From the training/validation dataset (D1, n=75 Z-stack images, 6 brain slices) and the independent testing dataset (D2, n=15 Z-stack images, 8 brain slices), we generated 3,000 synthetic training/validation images (SynD1) and 1,000 synthetic separate testing images (SynD2). We used SynD1 to train the Swin transformer-based deep learning network.

The deep neural network architecture

The architecture of the deep neural network based on the Swin transformer (Lin et al., 2021) is shown in Fig. 3. The neural network consists of four stages with four components: patch partition module, linear embedding, Swin transformer block, and patch merging module. The patch partition module separates the grayscale input images into non-overlapping 4×4 patches. In the NLP domain, tokenization is employed for cutting input text data into meaningful parts, called “tokens,” which can be embedded into a vector space and fed into the transformer. Like the NLP domain transformers, each patch is considered a token with a feature dimension of 4×4×1. Next, the linear embedding layer projects these features into an arbitrary dimension, C=96. This arbitrary dimension, C, determines the size of the Swin transformer. We used a small Swin transformer (C=96) for lower
computational complexity. The patches are then processed by multiple Swin transformer blocks, which preserve the number of patch tokens. The Swin transformer block has a shifted window-based multi-head self-attention module (SW-MSA) created by replacing the standard multi-head self-attention (MSA) (Dosovitskiy et al., 2020) with a shifted window-based module and a two-layer Multilayer Perceptron (MLP). The shifted window partitioning provides higher efficiency than standard multi-head self-attention by computing features within local windows and allowing cross-window connections. Before each SW-MSA and MLP, LayerNorm is applied, and a residual connection is made after each module. LayerNorm is a simpler normalization method that transforms the inputs to have zero mean and unit variance for each batch across all elements. The linear embedding module, followed by two consecutive Swin transformer blocks, is Stage 1.

Stage 2 consists of a patch merging module and two consecutive Swin transformer blocks. The patch merging layer merges the features from each group of 2×2 neighboring patches and applies a linear layer on the 4C dimensional concatenated features. The output from the patch merging layer has a dimension of 2C, with fewer patch tokens. Since the patch merging layer merges features from neighboring 2×2 patches, the number of tokens is reduced by a multiple of 2×2=4, and the resolution is downsampling by a factor of 2, resulting in $\frac{H}{8} \times \frac{W}{8}$. Next, two consecutive Swin transformer blocks are applied, with resolutions of $\frac{H}{8} \times \frac{W}{8}$. Stages 3 and 4 consist of similar modules and keep the resolutions at $\frac{H}{16} \times \frac{W}{16}$ and $\frac{H}{32} \times \frac{W}{32}$. The denominators represent further downsampling of resolutions of 2 by the patch merging layer. These four stages produce a hierarchical representation. Once trained, this deep learning network generates the output segmentations of the neural cell...
bodies from the input images with the corresponding confidence scores for each instance. We used thresholding to remove objects with lower confidence scores. The synthetic training/validation dataset (SynD1) was used for architecture design and training to determine the hyper-parameters of the Swin transformer. Using this architecture, the Swin transformer correctly identified neuronal boundaries in the independent testing dataset (SynD2; Fig. 4A). Even with salt and pepper noise, the Swin transformer detects and segments the neurons accurately (Fig. 4B).

**Swin transformer Performance**

We compared the performance of our Swin transformer to a Mask-R CNN-based algorithm (ANMAF) (Tong et al., 2021) and a CNN-based approach (Cellpose) (Pachitariu and Stringer, 2022; Stringer et al., 2021). We used GCaMP6s-expressing neurons imaged at baseline (aCSF perfusion) and NMDA perfusion/washout as conditions representing low and high fluorescence, respectively (n=5 brain slices in each condition). We evaluated the algorithms in the independent test dataset (D2, n=15 Z-stack images, 8 brain slices) and the independent synthetic test dataset (SynD2).

We first demonstrated the generality of the Swin transformer to segment maximum intensity projection (MIPs) images at different depths and in single-plane images. The model trained on 20 µm MIP images segments neurons using MIP images of different depths, in addition to single plane images (Fig. 5). The probability of false positives/doublets was higher with the increase in the depths of the MIPs images (ROI 1), however, the accuracy of the predicted boundary (ROI 2), and the total number of detected
neurons (ROI 3) was also higher with increased depths. We decided to use 20 µm depth MIPs as it balances accuracy with many neuronal detections, but this depth can be adjusted based on the researcher's preference.

We then determined the reliability of the Swin transformer for boundary detection by comparing the identified areas with the manually annotated area measurements. The average difference between the Swin transformer and manual areas was 34.2 ± 25.6 µm² (mean ± SD, n=113; Fig. 6). This overestimation of the Swin transformer (12.5 ± 9.3 %) could be explained by human annotators consciously avoiding proximal dendrite segments during manual tracing compared to the automated algorithm.

Next, we evaluated the effect of the detection threshold on the performance of the Swin transformer and compared it to ANMAF. When ANMAF was used to detect GCaMP6s-expressing neurons, whose brightness depends on intracellular Ca²⁺, it failed to detect and segment many neurons compared to the Swin transformer (Fig. 7, Table 2). Next, we varied the Swin detection threshold from 0.05 to 0.4 and analyzed its performance by examining 6 images (3 from baseline and 3 from NMDA). A threshold higher than 0.05 resulted in a higher percentage of correct detections and a lower percentage of incorrect detections. However, some true positives were also removed if the threshold value was too high, leading to fewer detections. In all conditions, all threshold values achieved a higher percentage of correct detections and a higher number of detections than ANMAF (Table 2).

When comparing the Swin transformer performance, we used a 0.05 threshold value to achieve the highest detections (highest recall) and minimize false negatives. This
parameter can be adjusted according to the researcher’s preference to achieve high recall (with more false positives) or high precision (with fewer false positives).

We next compared the performance of our Swin transformer to both ANMAF (Mask-R CNN-based algorithm) and Cellpose (a CNN-based approach, Fig. 8A). Here, the ground truth was represented by 131 neurons derived from the expert annotations. We computed the Sorensen-Dice coefficient from all predicted neurons for each ground truth instance of a neuron. The final identified neuron was the one that overlapped most closely with a ground truth neuron (highest Sorensen-Dice coefficient). An instance was labeled as correct if the ground truth neuron overlapped with a predicted neuron with an intersection-over-union (IOU) greater than 0.5. Otherwise, the instance was categorized as a missed neuron.

ANMAF correctly detected 118 of 131 neurons (90.1% yield rate). Cellpose also detected 118 of 131 neurons (90.1% yield rate). Instead, the Swin transformer had a higher detection yield of 98.5% (129 of 131 neurons, 0.05 threshold value; Table 3). There was a statistical difference in the Dice and IOU scores between the different algorithms. However, the Swin transformer’s Dice and IOU scores were not different from ANMAF or Cellpose (Fig. 8B, C). Therefore, the Swin transformer has a higher detection rate without losing accuracy in boundary demarcation.

Since the yield ratios were no different between ANMAF and Cellpose, we compared the performance of the Swin transformer to ANMAF. We computed the number of true positives (TP), false positives (FP), and false negatives (FN) for both approaches. Two human experts were provided with all the Swin transformer and ANMAF detections, who classified them as TP or FP. Additionally, they manually contoured the regions of the
neurons missed by the approaches to determine the number of false negatives. ANMAF detected 355 neurons, of which 314 were TP, 41 were FP, and 289 were missed neurons (Table 4). The Swin transformer detected 632 neurons, with 518 TP, 103 FP, and 11 instances labeled as “unknown” since the human experts could not categorize them as true or false positives (Table 4). The precision from ANMAF (0.88) was slightly higher than the Swin transformer approach (0.83). However, the Swin transformer approach had a much higher recall (0.86) than ANMAF (0.52; Table 4). Thus, the Swin transformer performs better than a Mask-R CNN algorithm (ANMAF) when neurons express GCaMP6s, with more detections, fewer false negatives, and more true positives.

We further evaluated if the Swin transformer detections capture biologically relevant changes to the neuronal somas during NMDA-induced cell swelling (30 µM NMDA for 10 mins) and compared it to ANMAF and manual annotations. All three approaches were used to detect neuronal somatic areas from 2-photon images of the same Thy1-GCaMP6s brain slices in three distinct conditions (baseline, NMDA, and washout, Fig. 9A). The manual somatic ROIs were hand-traced by one experimenter, and another blinded to the hand-traced ROIs evaluated the automated detections generated by ANMAF and the Swin transformer. Detections labeling the same neurons were paired across different detection methods and conditions based on the ROI spatial location. The Swin transformer neuronal detections had slightly larger areas than manually traced ROIs but not ANMAF-detected neurons (Fig. 9B). However, the Swin transformer generated more detections in all three conditions with greater accuracy than ANMAF (Table 5). Importantly, area overestimation by automated algorithms did not affect the estimation of the downstream physiological outcome, as the NMDA-induced prolonged neuronal swelling was captured in all three
detection methods to the same extent and not different between detection methods (Fig. 8C). Thus, changes to neuronal somatic areas are detected by the Swin transformer as precisely and carefully as hand tracing while eliminating human bias in sample selection. Furthermore, the efficacy of the Swin transformer to detect and segment neuronal somas in NMDA and washout conditions despite high neuropil signal (i.e., low contrast) demonstrates the effectiveness of the Swin transformer in the presence of noise.

Finally, we evaluated if the Swin transformer trained with a calcium biosensor (GCaMP6s) can detect neuronal cell bodies labeled with a stable fluorophore and in another brain region with a different cytoarchitecture. We imaged YFP-expressing neurons in the CA1 region of the dorsal hippocampus under two-photon microscopy. Neurons in the pyramidal layer in the hippocampal CA1 region are densely packed compared to the neocortical neurons used to train the Swin transformer. Despite this, the Swin transformer correctly marks cell bodies in YFP-expressing CA1 neurons even when trained with GCaMP6s labeled neurons (Fig. 10).

**Discussion**

It is crucial to accurately segment neuronal cell bodies to determine their changes during different physiological and pathological conditions. Despite the paramount importance, the segmentation of neuronal cell bodies is challenging due to their complex histology. Here, we developed an automated, accurate, reproducible, fast, and unbiased deep-learning-based technique capitalizing on the state-of-the-art Swin transformer (Lin et al., 2021) for neuronal cell body analysis, generalizable to low and high fluorescence conditions. Our algorithm detects and segments 2D somas of GCaMP6s and YFP-expressing neurons from
acute brain slices. The Swin transformer excelled in detecting even when trained with a few partially labeled neurons (approximately 700 manually annotated neurons), eliminating the need for time-consuming human annotations of many cells. It outperformed Mask-R CNN and CNN deep-learning techniques, correctly labeling more neurons. Our Swin transformer has flexibility in its implementation, adjusting to the researcher’s needs. It can detect neurons expressing a different fluorophore if they have a cytoarchitecture similar to the one it was trained for. Its threshold can be adjusted to balance recall and precision, and it can use different MIPs volumes.

Most previous approaches for neuronal cell body segmentation have used fixed, thin tissue slices and imaged them using confocal microscopy or transmitted electron microscopy (Cameron et al., 2020; Falk et al., 2019; Hosseini et al., 2020; Luengo-Sanchez et al., 2015). Also, prior approaches have been evaluated on neuronal cultures or thin brain tissue using a few neurons (Banerjee et al., 2020; Falk et al., 2019; Hosseini et al., 2020; Kayasandik and Labate, 2016; Li et al., 2019; Ozcan et al., 2015; Xu et al., 2016). However, thick brain slices (350 microns or more) commonly employed in physiological research have several challenges. These include overlapping neurons from different depth levels, changes in fluorescence brightness throughout the slice, or, in the case of GCamP6s, activity-dependent changes in fluorescence and low contrast between cell bodies and the background, among others. Current deep-learning-based techniques fail to detect and segment the boundaries accurately from thick brain slices, resulting in low recall and Dice scores whenever there is insufficient contrast between the cell and background or when there are many cells. Our Swin transformer approach achieves high detection and segmentation performance in those cases, providing a reliable, automated, fast, and
accurate tool for neuronal cell body analysis in thick brain slices. While the Swin transformer overestimated the areas by an average of 12.5% compared to manual tracing, this is systematic and not biased like human boundary labeling, which must decide where to limit the soma vs. the dendrite. It is important to note that while we used the Swin transformer to detect neuronal cell bodies using 20 µm maximum intensity projections, it can also be used with different depths of maximum intensity projections and even in a single plane (Fig. 5). For our purposes, we are interested in the maximal area of neuronal cell bodies (Glykys et al., 2019; Takezawa et al., 2023), which is why we use maximum intensity projections. While our work focuses on performing 2D segmentations, extending the approach for 3D segmentations may be possible by replacing the Swin transformer block in Fig. 3 with a Swin3D block (Yang et al., 2023). However, this would require exhaustive 3D slice-by-slice annotations of the neurons. Alternatively, the 2D segmentations can be performed on individual slices, and the predictions from the neighboring slices can be used to refine and generate 3D segmentation of neuronal volumes. Yet, this approach requires correlating neuronal boundaries between slices, which, if done manually, would take a significant amount of time. In the future, we plan to apply the current 2D model to segment the neurons on individual slices and then incorporate the information from the neighboring slices. This approach needs to be refined to yield accurate 3D segmentations.

The Swin transformer trained on GCaMP6s expressing neocortical neurons detected two-photon images of YFP-expressing pyramidal neurons in the CA1 region of the dorsal hippocampus. As such, this algorithm should be able to segment neurons labeled with other cytosolic fluorophores in other brain regions if neurons have a fluorescent cytoarchitecture.
similar to the trained data set. In case the current model does not perform to a user’s standard on a new fluorophore or brain region, more robust segmentation results can be obtained by using the current model as pre-trained weights to train a new model using only a few hundred manually annotated neurons of the target fluorophore, or brain region (transfer learning). Importantly, the Swin transformer detects meaningful increases in somatic neuronal areas similar to careful hand tracing, with no bias and quickly.

The Swin transformer approach has some limitations. It has a lower precision than ANMAF (0.83 vs. 0.88) due to higher predicted false positives. Some false positives are caused by poor contrast of the fluorescently labeled neurons with the background. Second, the Swin transformer measures neuronal somatic areas using maximum-intensity projections, not neuronal volume (3D segmentation). Third, the Swin transformer detects some doublets, where two adjacent neurons are identified as one instance. Human experts can avoid doublets by examining adjacent slices to obtain information about neuronal boundaries. In the case of the Swin transformer, the edges are predicted from only one image instead of using data from adjacent slices. We plan to address these limitations by developing a semi-supervised approach in the future to measure 3D neuronal volumes, which incorporates the prediction information from neighboring slices (Wang et al., 2022; Xu et al., 2021).

To conclude, we developed an automated deep-learning network based on a Swin transformer for 2D neuronal cell body analysis in thick brain slices. This tool can substantially impact neuroscience by detecting and accurately labeling numerous fluorescent neurons, even with a low-emitting fluorophore, leading to more comprehensive and quantifiable studies of neuronal cell bodies. With the help of our flexible algorithm,
researchers can study the changes in neuronal somas under various physiological and pathological conditions using large numbers of neurons without manually annotating them.
References


Hosseini SMH, Chen H, Jablonski MM (2020) Automatic detection and counting of retina cell nuclei using deep learning In: Medical Imaging 2020: Biomedical Applications in Molecular, Structural, and Functional Imaging , pp634–646.


Figure 1. Experiment design. Experimental design and acquisition time-points of multiphoton Z-stacks for the three conditions: during baseline aCSF, NMDA perfusion, and washout. An identical approach was used to acquire images from Thy1-YFP-expressing neurons.

Figure 2. Generation of synthetic neuronal soma images. A) Three examples of GCaMP6s-expressing neurons imaged under two-photon microscopy under three conditions: baseline aCSF, NMDA perfusion, and washout. The white rectangles represent the background, and the yellow outlines represent the soma boundaries traced by a human expert. B) Background extraction. (C) Synthetic background generation from the extracted tiles. The background tiles were shuffled and randomly placed to generate the synthetic background. D) Generation of synthetic neuronal images. The manually identified neuronal somatic boundaries were cut, pasted, and blended into the randomly selected background images to generate synthetic images. Scale bar= 50 μm.

Figure 3. Deep learning network architecture based on the Swin transformer. Two successive Swin transformer blocks are shown at the bottom. W-MSA and SW-MSA represent multi-head self-attention modules with regular and shifted windows. The orange outlines represent the output neuronal boundaries identified by the Swin transformer. Scale bar=50 μm.

Figure 4. The Swin transformer detected neuronal boundaries in the independent testing dataset (SynD2). A) Left column: synthetically generated images without noise. Right column: Swin transformer detected boundaries. B) Same as A in the presence of salt.
and pepper noise. Cyan outlines: True positive detections by the Swin transformer. Yellow outlines: false negatives.

**Figure 5. Neuronal segmentation using maximal intensity projections of different depths.** A) Maximal intensity projections (MIP) two-photon images z-stacks images of GCaMP6s expressing neocortical neurons imaged at 2 μm steps. Segmented neurons (magenta) are overlayed on each image. B) Three different ROIs from each image in panel A. Notice that the probability of false detection and doublets increases with overlapping neurons (ROI 1), the accuracy of correct predictions (ROI 2), and the total number of segmented neurons (ROI 3) increases with thicker MIPs. Images contrast-enhanced (CLAHE, ImageJ). Scale bar=50 μm.

**Figure 6. Neuronal somatic area comparison between the Swin transformer and manually detected ones.** The mean difference is 34.2 ± 25.6 μm² (SD). Slope 1.10 [95% CI 1.02-1.18]. Pearson correlation R²=0.875; p<0.001. N=113 somatic areas. Line: fit. Dashed lines: 95% confidence intervals.

**Figure 7. Improved efficacy of neuronal somatic boundary detection by the Swin transformer compared to a Mask-R CNN algorithm.** Representative images of GCaMP6s-expressing neurons under baseline aCSF perfusion (Left column) and NMDA perfusion (right column). A) ANMAF detected neurons. B) Swin transformer detected neurons. The correct detections are in cyan, missed detections in yellow, and false positive detections in magenta. Images contrast-enhanced (CLAHE, ImageJ). Scale bar=50 μm.

**Figure 8. Somatic detected boundaries by Swin transformer, Cellpose, and ANMAF show no difference.** A) Left column: baseline aCSF perfusion. Middle column: NMDA
perfusion. **Right column: washout.** Yellow outlines in the top row show the manually annotated neuronal boundaries. The ANMAF and Cellpose outputs are labeled with magenta and orange outlines in the second and third rows. The Swin transformer approach detected contours marked in cyan, third row, using a detection threshold value of 0.05. **B)** Dice scores, RM One-way ANOVA $F(2,212)=4.31$, $p=0.015$. Dunnett's multiple comparison tests, $p=0.224$ between Swin and ANMAF and $p=0.269$ between Swin and Cellpose. **C)** IOU scores, RM One-way ANOVA $F(2,212)=4.57$, $p=0.011$. Dunnett's multiple comparison tests, $p=0.248$ between Swin and ANMAF and $p=0.209$ between Swin and Cellpose. Means ± 95% CI. Images contrast-enhanced (CLAHE, ImageJ). Scale bar=50 μm.

**Figure 9. The Swin transformer detects changes in neuronal somatic sizes in the presence of NMDA.** A) Representative images showing ROIs obtained using hand tracing (yellow, top), ANMAF (magenta, middle), and Swin transformer (cyan, bottom). **B)** The Swin transformer calculated higher areas than manually labeled ROIs. **Baseline:** RM One-way ANOVA, $F(2,48)=8.72$, $p=0.0006$, Dunnett’s post-hoc test; $n=25$ matched neurons. **NMDA:** RM One-way ANOVA, $F(2,82)=8.695$, $p=0.0004$, Dunnett’s post-hoc test; $n=42$ matched neurons. **Washout:** RM One-way ANOVA, $F(2,56)=7.840$, $p=0.0010$, Dunnett’s post-hoc test; $n=29$ matched neurons. Two brain slices. **C)** Matched absolute somatic area change during NMDA-induced long-term neuronal swelling across three detection methods. Two-way ANOVA: Interaction $F(4,141)=0.32$, $p=0.86$; Treatment $F(2,141)=20.33$, $p<0.0001$; Detection algorithm $F(2,141)=2.67$, $p=0.073$. **D)** Matched area percent change during NMDA-induced long-term neuronal swelling across three detection methods. Two-way ANOVA: Interaction $F(4,141)=0.4295$, $p=0.787$; Treatment $F(2,141)=22.88$, $p<0.0001$; Detection algorithm $F(2,141)=1.381$, $p=0.255$. Manual $n=13$ matched neurons across 29

**Figure 10. Neuronal boundaries of YFP-expressing neurons in the CA1 region detected by Swin transformer.** The Swin transformer was trained with GCaMP6s labeled neurons. Baseline aCSF perfusion (Left column) and, NMDA perfusion (middle column), and Washout perfusion (right column). Images contrast-enhanced (CLAHE, ImageJ). Scale bar = 50 μm

<table>
<thead>
<tr>
<th>Figure</th>
<th>Graph</th>
<th>Structure</th>
<th>Type of test</th>
<th>Test result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fig. 7</td>
<td>B</td>
<td>Normal</td>
<td>RM One-way ANOVA</td>
<td>F(2, 212)=4.31, p=0.015 Dunnett’s:  p=0.224 (Swin/ANMAF)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Dunnett’s:  p=0.269 (Swin/Cellpose)</td>
</tr>
<tr>
<td>Fig. 7</td>
<td>C</td>
<td>Normal</td>
<td>RM One-way ANOVA</td>
<td>F(2, 212)=4.57, p=0.011 Dunnett’s:  p=0.209 (Swin/ANMAF)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Dunnett’s:  p=0.248 (Swin/Cellpose)</td>
</tr>
<tr>
<td>Fig. 8</td>
<td>B (Base)</td>
<td>Normal</td>
<td>RM One-way ANOVA</td>
<td>F(2,48)=8.72, p=0.0006 Dunnett’s:  p=0.0004 (Swin/Manual)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Dunnett’s:  p=0.5230 (Swin/ANMAF)</td>
</tr>
<tr>
<td>Fig. 8</td>
<td>B (NMDA)</td>
<td>Normal</td>
<td>RM One-way ANOVA</td>
<td>F(2,82)=8.695, p=0.0004 Dunnett’s:  p=0.0004 (Swin/Manual)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Dunnett’s:  p=0.6743 (Swin/ANMAF)</td>
</tr>
<tr>
<td>Fig. 8</td>
<td>B (Wash)</td>
<td>Normal</td>
<td>RM One-way ANOVA</td>
<td>F(2,56)=7.840, p=0.0010 Dunnett’s:  p=0.0011 (Swin/Manual)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Dunnett’s:  p=0.8091 (Swin/ANMAF)</td>
</tr>
<tr>
<td>Fig. 8</td>
<td>C</td>
<td>Normal</td>
<td>Two-way ANOVA</td>
<td>Interaction F(4,141)=0.32, p=0.86</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Treatment F(2,141)=20.33, p&lt;0.0001 Detection algorithm F(2,141)=2.67, p=0.073</td>
</tr>
<tr>
<td>Fig. 8</td>
<td>D</td>
<td>Normal</td>
<td>Two-way ANOVA</td>
<td>Interaction F(4,141)=0.4295, p=0.787</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Treatment F(2,141)=22.88, p&lt;0.0001 Detection algorithm F(2,141)=1.381, p=0.255</td>
</tr>
</tbody>
</table>

**Table 1. Statistical Table.**
Table 2. Threshold parameter changes on the Swin transformer performance. N=3 Z-stack images for the Baseline and three for the NMDA condition. FP: false positives.

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Baseline</th>
<th>NMADA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total detected neurons</td>
<td>Correct detections</td>
</tr>
<tr>
<td>ANMAF</td>
<td>54</td>
<td>55.6 %</td>
</tr>
<tr>
<td>Swin 0.05</td>
<td>105</td>
<td>66.7 %</td>
</tr>
<tr>
<td>Swin 0.10</td>
<td>98</td>
<td>76.5 %</td>
</tr>
<tr>
<td>Swin 0.20</td>
<td>92</td>
<td>77.2 %</td>
</tr>
<tr>
<td>Swin 0.40</td>
<td>82</td>
<td>81.7 %</td>
</tr>
</tbody>
</table>

Table 3. Quantitative comparisons between the ANMAF, Cellpose, and the Swin transformer on the manually annotated neurons from the independent testing dataset (D2). N=15 Z-stack images, 8 brain slices.
Table 4. Comparison between the approaches for detecting neurons on the independent testing dataset (D2). N=15 Z-stack images, 8 brain slices.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Total identified neurons</th>
<th>Correct detections</th>
<th>Wrong boundary</th>
<th>Total identified neurons</th>
<th>Correct detections</th>
<th>Wrong boundary</th>
<th>Total identified neurons</th>
<th>Correct detections</th>
<th>Wrong boundary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manual</td>
<td>48</td>
<td>-</td>
<td>-</td>
<td>56</td>
<td>-</td>
<td>-</td>
<td>44</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ANMAF</td>
<td>52</td>
<td>78.9 %</td>
<td>21.1 %</td>
<td>101</td>
<td>74.3 %</td>
<td>25.7 %</td>
<td>91</td>
<td>64.8 %</td>
<td>35.2 %</td>
</tr>
<tr>
<td>Swin</td>
<td>70</td>
<td>82.9 %</td>
<td>17.1 %</td>
<td>140</td>
<td>75.7 %</td>
<td>24.3 %</td>
<td>114</td>
<td>78.1 %</td>
<td>21.9 %</td>
</tr>
</tbody>
</table>

Table 5. Accuracy of neuronal morphology detections using hand tracing, ANMAF, and the Swin transformer. N= 6 Z-stack images (2 images per condition), 2 brain slices.

Extended Data 1: Swin Transformer code
Acute slices (p8-12), Thy1-GCaMP6s → 2P imaging → NMDA (30 µM) treatment

Baseline 0' 10' 20' 30' 40' Time (min) → Washout
A

aCSF  NMDA  Washout

Real images

B

Extraction of tiles

C

Synthetic backgrounds

D

Generated synthetic images