Self-supervised Learning for Segmentation and Quantification of Dopamine Neurons in Parkinson's Disease

Abstract

Parkinson's Disease (PD) is the second most common neurodegenerative disease in humans. PD is characterized by the gradual loss of dopaminergic neurons in the Substantia Nigra (SN, a part of the mid-brain). Counting the number of dopaminergic neurons in the SN is one of the most important indexes in evaluating drug efficacy in PD animal models. Currently, analyzing and quantifying dopaminergic neurons is conducted manually by experts through analysis of digital pathology images which is laborious, time-consuming, and highly subjective. As such, a reliable and unbiased automated system is demanded for the quantification of dopaminergic neurons in digital pathology images. We propose an end-to-end deep learning framework for the segmentation and quantification of dopaminergic neurons in PD animal models. To the best of knowledge, this is the first machine learning model that detects the cell body of dopaminergic neurons, counts the number of dopaminergic neurons and provides characteristics of individual dopaminergic neurons as a numerical output. Extensive experiments demonstrate the effectiveness of our model in quantifying neurons with a high precision, which can provide quicker turnaround for drug efficacy studies, better understanding of dopaminergic neuronal health status and unbiased results in PD pre-clinical research.

1 Introduction

Image segmentation is a fundamental tool to develop artificial intelligence medical imaging platform [1], specifically in the area of radiology and digital pathology. For instance, deep learning cell segmentation models can enable robust and fast approaches to quantify cells in histopathology images, enabling more sensitive analysis of biological experiments in animals and humans [2, 3]. However, deep learning models rely on large-scale high quality data, limiting their applications in biological use cases. In this manuscript, we implement self-supervised learning techniques to develop a robust automated pipeline to segment specialized neuronal cell known as dopaminergic neurons (DA neurons) and quantify them to examine DA neuronal loss in preclinical studies and understand Parkinson's disease (PD) biology.

In this study, a deep learning-based framework segments and detects Tyrosine Hydroxylase (TH) positive DA neurons in the SN of mouse brain tissues. SN is the area of the mid-brain that consists of DA neurons which are most susceptible to genetic and sporadic factors and lost in Parkinson's disease (PD). TH is an enzyme that is specifically expressed in DA neurons. TH staining is the most reliable method used for detecting DA neurons. TH stains the soma (cell body), nucleus and the axons of DA neurons. Loss of dopaminergic neurons leads to motor neuron associated dysfunctions as observed in PD patients and animal models [4]. Preventing loss of DA neurons is the most important goal of PD targeting therapies. The TH staining intensity is also an indicator of the health status of the DA neurons and is considered the most reliable method to identify loss of DA neurons [5]. Pre-clinical research on PD is highly dependent on segmentation and quantification of DA neurons in the SN [6, 7]. The unique morphology and organization of DA neurons also makes it difficult to use generalized cell segmenting models to identify them and delve deeper into understanding it's biology. The challenges with the current DA neuron analyzing strategies are: traditional stereology is extremely time consuming and subjective to user associated bias; Threshold for optical density based method cannot capture the complex nature of DA neurons on histology sections and

provide inaccurate numbers; Nucleus segmentation method also fail to detect individual cells and cannot retrieve the information on cell body (i.e. TH intensity in soma) which can be used to interpret the biology of DA neurons. Generalist cell segmentation automated model such as Cellpose [8] have been developed to solve this problem but its efficiency in detecting specific type of neurons such as DA neurons is still limited [9]. A generalist model does not provide additional information that is specific to DA neurons which holds high value in PD research. Hence, it has become crucial to develop a machine learning model that can analyze and quantify DA neurons precisely in the SN with a quick turnaround time and immune to user associated bias. This will in turn make a huge impact in the field of PD pre-clinical research by identifying the efficacy of potent drugs in a shorter time-frame and accelerating the possibility of taking a potential drug into the clinic.

Our model leverages a combination of data sampling techniques and cross-domain self supervised learning [10, 11] on both unlabeled natural images and domain specific pathology images to learn transferable and generalize representations for pathology images. Such representations can be further fine-tuned and deployed for the neuronal cell segmentation using limited labeled data from the biological experiments. We compare the performance of fine tuned model which is originally trained on different data, (1) natural images, (2) pathology images, or (3) natural images followed by digital pathology images. We next compare the predicted number of TH cells from our model to manual counts done by histopathology experts to investigate the accuracy of automated quantification. Furthermore, we analyze the effects of the combination of various augmentation methods on the segmentation performance of the model. Experimental results and extensive analysis indicate that our model can outperform existing models, especially in low data scenarios.

In summary, we make the following contributions:

- The first end-to-end framework for automatic segmentation and quantification of DA neurons in whole-slide digital pathology images of PD models.
- A cross-domain self-supervised pre-training approach that exploits the power of unlabeled natural and medical images for representation learning.
- A comprehensive set of experiments that demonstrate the effectiveness and efficiency of our model in detecting and quantifying DA neurons using a limited amount of annotated data.
- A numerical and visual data output to indicate the phenotypic characteristics of DA neurons segmented by the model

2 Related Works

CNN-based quantification of dopaminergic neurons. Deep learning methods have been successfully utilized in analyzing human digital pathology images for different tasks, including cell segmentation and cell counting [12, 13, 3, 2, 14]. Generalist cell segmentation models such as Cellpose [8] have been developed to segment many types of cell. Cellpose relies on a large dataset of images of different cells and a reversible transformation from training set segmentation masks to vector gradients that can be predicted by a neural network. In particular, it leverages a U-Net model to predict the horizontal and vertical gradients, as well as whether a pixel belongs to any cell. The three predicted maps are then combined into a gradient vector field. The predicted gradient vector fields are used to construct a dynamical system with fixed points whose basins of attraction represent the predicted masks. Despite its success, the efficiency of Cellpose in detecting specific type of neurons such as DA neurons is still limited [9]. The number of studies that employ deep learning for the quantization of DA neurons in animal models of PD are relatively limited. [15] Penttinen et al implemented a deep learning-based method for processing whole-slide digital imaging to count DA neurons in SN of rat and mouse models. This study leverages the TH positive nucleus to detect the TH cells which is susceptible to error because of the architecture of DA neurons in SN and makes it difficult to distinguish between overlapping cells when detected only relying on nucleus as annotations. [16] In Zhao et al study, they developed a framework for automatic localization of SN region and detection of neurons within this

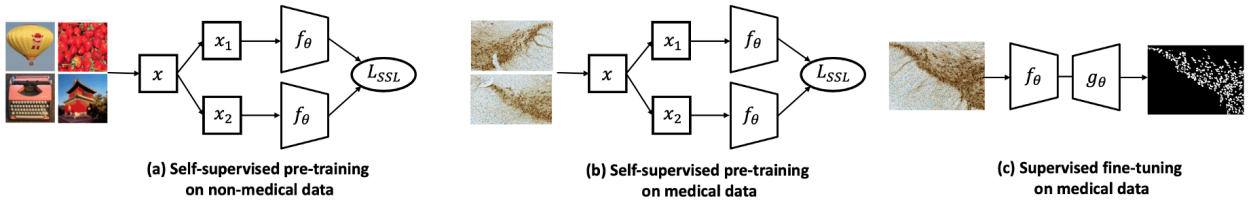


Figure 1: An overview of our approach. To address the annotated data scarcity challenge for training deep models, we perform (a) self-supervised pre-training on natural images, and then (b) self-supervised pre-training on digital pathology images. We finally (c) fine-tune the self-supervised pre-trained model with limited annotated data for target neuron segmentation task.

region. The SN localization is achieved by using a Faster-RCNN network, whereas neuron detection is done using a LSTM network. However, these studies are limited to counting neurons and/or detecting neuron locations and do not provide additional information about individual cells, such as TH intensity and health status, which is essential for understanding the biology behind DA neuronal loss and its association with PD pathogenesis.

Self-supervised Learning. Self-supervised learning methods aim to learn generalizable representations from unlabeled data. This paradigm involves training a neural network on a manually created (pretext) task for which ground truth is obtained from the data. The learned representations can be transferred and fine-tuned for various target tasks with limited labeled data. Instance discrimination methods [17, 18, 19, 20, 21, 22] have recently sparked a renaissance in the SSL paradigm. These methods consider each image as a separate class and seek to learn representations that are invariant to image distortions. Contrastive approaches, such as MoCo [18], consider the augmented views of the same image as positive pairs and the ones from the other images as negative pairs, and aim to enhance the similarity between positive pairs while decreasing the similarity between negative pairs. Barlow Twins [17] minimizes redundancy in the representations by measuring the cross-correlation matrix between the embeddings of two views of the same image and making it close to the identity matrix. Clustering approaches, such as deep cluster [19] and SwAV [19], simultaneously cluster the data while enforcing agreement between cluster assignments generated for different views of the same image. Motivated by the success in computer vision, instance discrimination SSL methods have been adopted in medical applications. A recent transfer learning study for medical imaging [23] demonstrated the efficacy of existing instance discrimination methods pre-trained on ImageNet for various medical tasks. A group of work focused on designing SSL frameworks by exploiting consistent anatomical structure within radiology scans [24, 25]. Another line of studies designed contrastive-based SSL for medical tasks [11, 26, 27, 10], including whole slide image classification [28]. In contrast to the previous works, our work is the first study that investigates the efficacy of SSL for digital pathology images of PD animal models to compensate the lack of large-scale annotated datasets for training deep learning models.

3 Method

3.1 Animal studies, annotations, and dataset

The data-set used in this study was obtained by manually labeling 30,000 TH positive DA neuronal soma in 2D histology digital images. This is an internal data-set. The digital images were obtained from multiple animal studies where mouse brains were sectioned at 35 micron thickness and stained with TH and either Haematoxylin or Nissl as a background tissue stain. Each animal study consisted of multiple animals and all the animals in one study were stained at one time-frame. This allowed us to take into account the minor tissue processing and staining associated variability that commonly occurs in histology studies. The sections were then imaged using a whole slide scanner microscope, Nanozoomer system (Hamamatsu Corp, San Jose,

CA) at 20x resolution (0.46 microns/pixel). Whole coronal brain section images containing the SN were exported from the digital scans at 20x resolution and were used to annotate the TH positive DA neuronal soma and train the model. The ground truth (GT) for this study was labelled and quality controlled by biologists who specialize in mouse brain anatomy and PD research. We have evaluated human annotator bias in one of our previous neuroanatomy segmentation studies that have been published. We observed a 5-8 percent difference in output with different manual annotators. Since manual annotation and counting of individual cells is extremely laborious in a data-set used in this study (our data-set had approximately 30,000 individual cells), to mitigate the human annotator bias, we had 3 annotators who randomly annotated the cells for the GT in this study. A neuroanatomy expert further QCd the annotations and made the necessary changes to reflect the cell annotation. The representative image is depicted in Fig 2 ground truth panel. The blind test data-set (out-study test set) used for analyzing model's efficiency was a separate animal study in which the model has not been directly trained on the study group. This data-set consists of 20 brain sections randomly selected from a different animal study with 150 brain sections. The ground truth (GT) for the out-study test set was established by annotating the soma. Our model and Cellpose were tested on these 20 sections, and the results are shown in Figures 4-6. The DA neurons were detected by the model (red) and visually represented to compare it with the manually counted neurons (blue) by the biologist.

3.2 Self-supervised Pre-training

Our approach is established on continual self-supervised pre-training in which a model is first pre-trained on a massive general dataset, such as ImageNet, and then pre-trained on domain-specific datasets. For the first step (see Figure 1.a), we train the self-supervised model on the ImageNet dataset using state-of-the-art instance discrimination approaches, such as Barlow Twins [17]. For the second step (see Figure 1.b), we continue the self-supervised pre-training on the in-domain medical dataset. Finally, we fine-tune the pre-trained models for the neuron segmentation (target) task using labeled images (see Figure 1.c).

Barlow Twins [17]. This SSL approach aims to reduce the amount of redundant information about each sample in the learnt representations while simultaneously making the representation invariance to image distortions. To do so, given an image sample X, two distorted views of the sample are generated by applying a data augmentation function $\mathcal{T}(.)$ on X. The two distorted views X_1 and X_2 are then processed by the backbone network f_θ to produce latent representations $Z_1 = f_\theta(\mathcal{T}(X_1))$ and $Z_2 = f_\theta(\mathcal{T}(X_2))$. The backbone network f_θ includes a standard ResNet-50 encoder and a three-layer MLP projection head. The model is trained by minimizing the following loss function:

$$\mathcal{L}_{SSL} = \sum_i (1 - \mathcal{C}_{ii})^2 + \lambda \sum_i \sum_{i \neq j} \mathcal{C}_{ij}^2 \tag{1}$$

where $\mathcal{C}$ is the cross-correlation matrix computed between Z_1 and Z_2 along the batch dimension. λ is a coefficient to identify the weight of each loss terms. The model is trained by making the cross-correlation matrix $\mathcal{C}$ close to the identity matrix. In particular, by equating the diagonal elements of the $\mathcal{C}$ to 1, the learned representation will be invariant to the image distortions. By equating the off-diagonal elements of the $\mathcal{C}$ to 0, the different elements of the representation will be decorrelated, so that the output units contain non-redundant information about the images.

3.3 Network Architecture.

For target segmentation task, we use a U-Net network which consists of encoder (f_θ) and decoder (g_θ) parts. The encoder is a standard ResNet-50, which is initialized with the self-supervised pre-trained encoder.

3.4 Tile sampling and augmentation.

For target segmentation task, we divide images into non-overlapping patches of size 512×512 to ensure we sample from every part of the image. In all experiments, the raw image intensities per channel are

Pre-training	Initialization	Dice(%)
-	Random	86.43±0.96
Supervised	ImageNet	85.86±3.37
Self-supervised	DeepCluster-v2	87.13±0.69
	Barlow Twins	87.24±0.75
	SwAV	**87.73±0.68**

(a) Fine-tuning with 100% of data: the best self-supervised method (i.e SwAV) yields significant boost ($p < 0.05$) compared with the best baseline (i.e. training from random initialization).

Pre-training	Initialization	Dice(%)
-	Random	67.22±8.24
Supervised	ImageNet	76.76±4.25
Self-supervised	DeepCluster-v2	78.72±3.98
	Barlow Twins	79.50±2.02
	SwAV	**80.83±1.17**

(b) Fine-tuning with 25% of data: the best self-supervised method (i.e SwAV) yields significant boost ($p < 0.05$) compared with the best baseline (i.e. supervised ImageNet model).

Table 1: Comparison of different initialization methods on target segmentation task.

normalized to the [0,1]. Data augmentation is essential for biological and medical image analysis due to the typically limited amount of available annotated data. We use different data augmentation techniques to enforce the model to capture more robust and generalizable representations. In particular, we use Flip, Rotation, RGBShift, Blur, GaussianNoise, and RandomResizedCrop to teach the expected appearance and color variation to the deep model.

3.5 Fine-tuning protocol

. We initialize the encoder of the target model (i.e. U-Net) with the pre-trained models and fine-tune all target model parameters. We train the target models using the Adam optimizer with a learning rate of 1e-3 and (β_1, β_2) = (0.9,0.999). We use ReduceLROnPlateau learning rate decay scheduler. We use batch size of 32 and train all models for 200 epochs. We employ early-stop mechanism using the validation data to avoid over-fitting. We use Dice coefficient loss function for training the target task. Dice coefficient is used for evaluating the accuracy of the target segmentation task. We run each method ten times on downstream task and report the average and standard deviation performance over all runs.

3.6 Cell counting.

A naive method for automatic counting of cells from segmentation predictions involves computing the connected components within the masks and considering the number of connected components as the cell count. However, this approach may not be accurate due to the presence of overlapping cells that share boundaries. To address this issue and improve counting accuracy, we first calculate the minimum and average cell size using the ground truth for the training data. Then, we take the models predictions (segmentation masks) and extract the connected components within the prediction masks; each connected component represents one or more cells (in the case of overlapping cells). We then filter out components that are smaller than the minimum cell size. For the remaining components, we count cells by dividing the cell size by the average cell size.

4 Experiments and Results

Self-supervised Models Provide More Generalizable Representations

Experimental setup. In this experiment, we evaluate the transferability of three popular SSL methods using officially released models, including DeepCluster-v2 [19], Barlow Twins [17], and SwAV [19]. All SSL models are pre-trained on the ImageNet dataset and employ a ResNet-50 backbone. As the baseline, we consider (1) training the target model from random initialization (without pre-training) and (2) transfer learning from the standard supervised pre-trained model on ImageNet, which is the *de facto* transfer learning pipeline in medical imaging [11]. Both baselines benefit from the same ResNet-50 backbone as the SSL models.

Pre-training Method	Pre-training Dataset	Dice(%)
Random	-	67.22±8.24
Barlow Twins	ImageNet	79.50±2.02
SwAV	ImageNet	80.83±1.17
Barlow Twins	In-domain	70.92±5.41
Barlow Twins	ImageNet→In-domain	**81.73±1.03**

Table 2: Comparison of pre-training dataset for self-supervised learning.

Results. Table 1a displays the results, from which we draw the following conclusions: (1) transfer learning from the supervised ImageNet model lags behind training from random initialization. We attribute this inferior performance to the remarkable domain shift between the pre-training and target tasks. In particular, supervised ImageNet models are encouraged to capture domain-specific semantic features, which may be inefficient when the pre-training and target data distributions are far apart. Our observation is in line with recent studies [29] on different medical tasks suggests that transfer learning from supervised ImageNet models may offer limited performance gains when the target dataset scale is large enough to compensate for the lack of pre-training. (2) Transfer learning from self-supervised models provide superior performance compared with both training from random initialization and transfer learning from the supervised ImageNet model. In particular, the best self-supervised model (i.e. SwAV) yields 1.3% and 2.27% performance boosts compared with training from random initialization and the supervised ImageNet model, respectively. Our statistical analysis based on independent two-sample t-test demonstrates the significance ($p < 0.05$) of the gain provided by SwAV compared with random initialization. Intuitively, self-supervised pre-trained models, in contrast to supervised pre-trained models, encode features that are not biased to task-relevant semantics, providing improvement across domains. Our observation in accordance with previous studies [23] demonstrates the effectiveness of self-supervised ImageNet models for medical applications.

Self-supervised Models Provide Superior Performance in Semi-supervised Learning
Experimental setup. We conduct further experiments to evaluate the advantage that self-supervised pre-trained models can provide for small data regimes. To do so, we randomly select 25% of the training data and fine-tune the self-supervised pre-trained models on this subset of data. We then compare the performance of self-supervised models with training the target model from random initialization and fine-tuning the supervised ImageNet model.
Results. The results are shown in Table 1b. First, we observe that transfer learning from either supervised or self-supervised pre-trained models offer significant ($p < 0.05$) performance improvements compared with training from random initialization. In particular, the supervised ImageNet model provides a 9.5% performance improvement compared to the random initialization of the target model. Moreover, self-supervised models– DeepCluster-v2, Barlow Twins, and SwAV, offer 11.5%, 12.3%, and 13.6% performance boosts, respectively, in comparison with random initialization. These observations imply the effectiveness of pre-training in providing more robust target models in low data regimes. Second, we observe that self-supervised models provide significantly better performance than the supervised ImageNet model. Specifically, DeepCluster-v2, Barlow Twins, and SwAV achieve 1.96%, 2.74%, and 4% performance boosts, respectively, compared to the supervised ImageNet baseline. Our statistical analysis based on independent two-sample t-test demonstrates the significance ($p < 0.05$) of the gains provided by the best self-supervised models (i.e. SwAV and Barlow Twins) compared with the supervised ImageNet model. These observations restate the efficacy of self-supervised models in delivering more generic representations that can be used for target tasks with limited data, resulting in reduced annotation costs.

Impact of Pre-training Data on Self-Supervised Learning
Experimental setup. We investigate the impact of pre-training datasets on self-supervised learning. To do so, we train Barlow Twins on three data schemas, including (1) SSL on the ImageNet dataset, (2) SSL on

Metric	Score (%)
Precision	95.25
Recall	95.49
F1-score	95.31

(a) The results for counting precision, recall and F1-score of our method vs. human observers.

Method	Counting Error (%)
Connected components	21.66
Our approach	**9.08**

(b) The results for automatic neuron counting error compared with human counting.

Table 3: Neuron detection and counting results

the medical dataset (referred to as the in-domain), and (3) SSL on both ImageNet and in-domain datasets (referred to as ImageNet→In-domain). For ImageNet→In-domain pre-training, we initialize the model with SvAW pre-trained on ImageNet, followed by SSL on our in-domain dataset. We fine-tune all pre-trained models for the neuron segmentation task using 25% of training data.

Results. Table 2 shows the segmentation accuracy measured by the Dice score (%) for different pretraining scenarios. First, we observe that pre-training on only in-domain dataset yields lower performance than pre-training on only the ImageNet dataset. We attribute this inferior performance to the limited number of in-domain pre-training data compared with the ImageNet dataset (1500 vs. 1.3M). Moreover, we observe that the best performance is achieved when both ImageNet and in-domain datasets are utilized for pre-training. In particular, ImageNet→In-domain pre-training surpasses both in-domain and ImageNet pre-trained models. Our statistical analysis reveals that pretraining on ImageNet followed by pretraining on In-domain data provides significant boosts ($p < 0.05$) compared with both in-domain pretraining and ImageNet pretraining standalone. These results imply that pre-training on ImageNet is complementary to pre-training on in-domain medical datasets, resulting in more powerful representations for medical applications.

Dopaminergic Neuron Detection and Counting

Experimental setup. The DA neurons segmented by the model were compared to the DA neurons detected by a biologist in the same tissue section from the blind data-set. The biologist detected the DA neurons and counted them manually on an image analysis platform ImageJ. The output from the model was overlaid with the manually detected cells and based on the color coding of the DA neurons by the model, the true positive (TP), false positive (FP) and false negative (FN) were calculated by the biologist. We calculated precision, recall and F1-score metrics for the detected neurons in the test images. In these measures, TP is the number of neurons successfully detected by the model; FP is the number of neurons detected by the model but are not actually neurons; and FN is the number of neurons not detected by the model. We further compare the performance of our method in neuron counting to human counting. To do so, we calculate the percentage error between the total number of neurons counted by our method and human counting. We also conduct an ablation study to illustrate the superiority of our cell counting method over the naive approach of counting cells by the number of connected components in the images.

Results. The performance metrics for neuron detection are shown in Table 3a. As seen, our method can effectively detect dopaminerginc neurons in whole-slide digital pathology images; in particular, our approach achieves a precision, recall, and F1-score of 95.25%, 95.49%, and 95.31%, respectively. Moreover, Table 3b presents the neuron counting results against human counting. As seen, automatic counting of the cells through computing the connected components within segmentation masks yields an error rate of 21.66%, while incorporating the connected components' sizes in counting significantly decreases the error rate to 9%. This results demonstrate the effectiveness of our approach in handling the overlapping neurons and providing a reliable automatic system for neuron counting.

Ablation Experiments

Experimental setup. We conduct extensive ablation experiments on different data augmentation techniques and network architectures. We examine seven different combinations of transformation that are

Data Augmentation	Flip	Rotation	RandomBrightnessContrast	RandomGamma	RGBShift	Blur	GaussianNoise	RandomResizedCrop	Elastic Transformatio	Dice(%)
mode1	-	-	-	-	-	-	-	-	-	78.96±1.85
mode2	✓	✓	✓	✓	-	-	-	-	-	80.83±1.17
mode3	✓	✓	-	-	✓	✓	✓	-	-	80.64±1.06
mode4	✓	✓	-	-	✓	✓	✓	✓	-	**81.94±0.74**
mode5	✓	✓	-	-	✓	✓	✓	✓	✓	79.97±2.73
mode6	✓	✓	✓	✓	✓	✓	✓	✓	-	81.30±0.93
mode7	✓	✓	✓	✓	✓	✓	✓	✓	✓	80.43±1.18

(a) Comparison of different data augmentations.

Network Architecture	Dice(%)
DeepLabV3+	81.53±0.76
U-Net	**81.94±0.74**

(b) Comparison of different network architectures.

Table 4: Ablation Experiments.

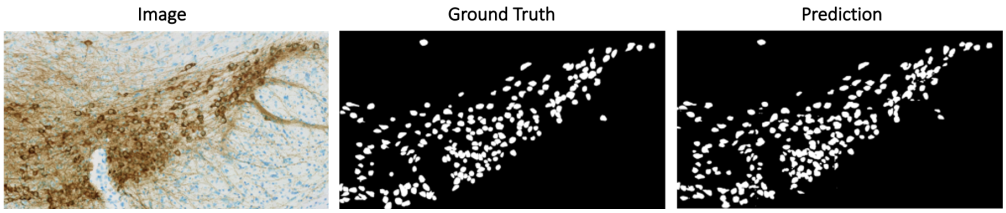


Figure 2: Visualization of Mouse brain 2D Image depicting DA neurons in the SN and segmentation results produced by our method.

commonly used in the literature, including (1) no augmentation (mode 1), (2) Flip, Rotation, RandomBrightnessContrast, and RandomGamma (mode 2), (3) Flip, Rotation, RGBShift, Blur, GaussianNoise (mode 3), (4) Flip, Rotation, RGBShift, Blur, GaussianNoise, RandomResizedCrop (mode 4), (5) Flip, Rotation, RGBShift, Blur, GaussianNoise, RandomResizedCrop, Elastic Transformation (mode 5), (6) Flip, Rotation, RandomBrightnessContrast, RandomGamma, RGBShift, Blur, GaussianNoise, RandomResizedCrop (mode 6), and (7) Flip, Rotation, RandomBrightnessContrast, RandomGamma, RGBShift, Blur, GaussianNoise, RandomResizedCrop, Elastic Transformation (mode 7). For network architectures, we examine U-Net and DeepLabV3+. In ablation experiments, all models are initialized with SvAW pre-trained model and fine-tuned with 25% of data.

Results. Table 4a shows the results of different data augmentation techniques. According to these results, the lowest performance comes from mode 1 (no augmentation), highlighting that combining pre-training with data augmentation techniques yields more accurate segmentation results for downstream tasks with limited amounts of data. Additionally, the combination of Flip, Rotation, RGBShift, Blur, GaussianNoise, RandomResizedCrop (mode 4) provides the best performance among all data augmentation approaches. This implies that color transformations such as RGBShift, Blur, and GaussianNoise can help the deep model in gleaning more generalizable representations. Furthermore, a comparison of the results obtained by modes 3 and 4, the latter of which includes an additional RandomResizedCrop, reveals that random cropping significantly contributes to performance improvements. Moreover, a comparison of the results obtained

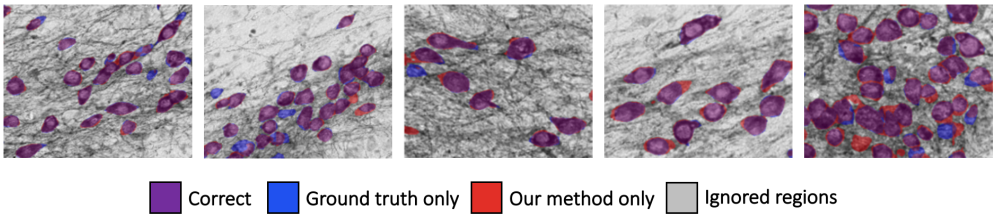
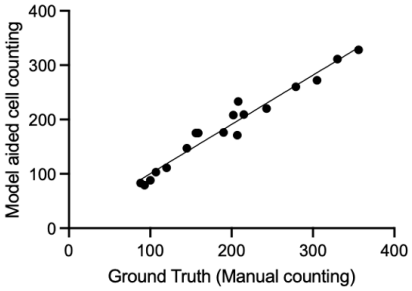


Figure 3: Visualization of cell segmentation results.



	Ground Truth (Manual counting) vs. Model aided counting
Pearson r	
r	0.9793
95% confidence interval	0.9442 to 0.9924
R squared	0.9591
P value	
P (two-tailed)	<0.0001
P value summary	****
Significant? (alpha = 0.05)	Yes
Number of XY Pairs	18

Figure 4: Correlation plot depicting the number of DA neurons counted by a biologist vs the number of DA neurons counted by the model. Blind data set was used to count the neurons from 18 brain sections stained with TH staining to identify the DA neurons and Nissl stain to stain the brain tissue. The sections for this study were chosen from multiple animal studies.

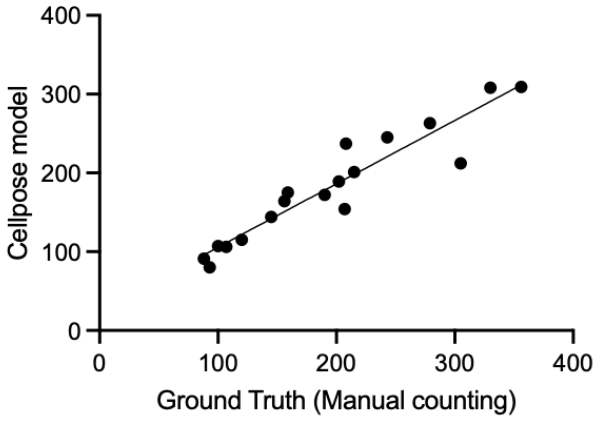
by modes 4 and 5, the latter of which includes an additional elastic transformation, demonstrates that elastic transformation has a negative impact on performance; the same observation can be drawn from the comparison of modes 6 and 7.

Table 4b presents the results of different network architectures for downstream neuron segmentation task. As seen, U-Net, which was originally designed for medical segmentation tasks, provides superior performance over DeepLabV3+.

Qualitative Results

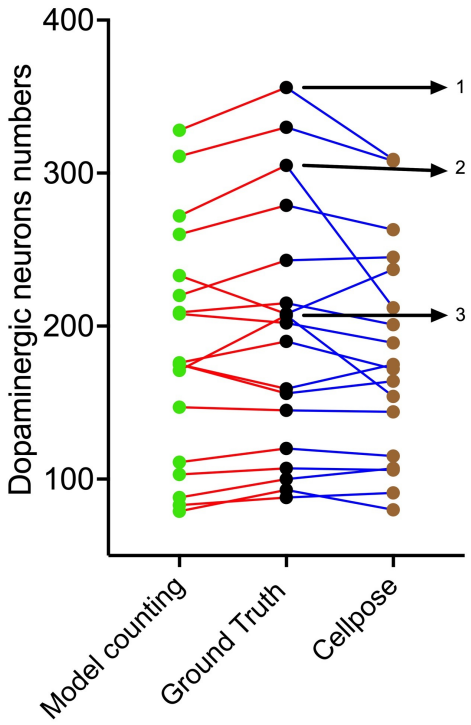
Experimental setup. We visualize the segmentation results of our best model from Table 1 on the test data. To do so, we first employ zero padding to make the size of the test images equal to a power of 512. Then, we divide the test images into non-overlapping 512×512 patches and then feed patches to the network. We then assemble the model's predictions for images patches to generate the prediction for the whole image. To examine the model's efficiency in counting DA neurons, a biologist counted the cells manually (Ground Truth) in the same section (blind dataset). We then ran a correlation statistics to measure the R^2 between the model and the GT. We additionally compared the GT to the latest generalist cell segmentation model-Cellpose and ran a correlation statistics to compare. Finally, the counts for DA neurons from our model, GT and Cellpose were plotted head to head to examine the efficiency of our model.We measured the TH intensity after converting the image into grayscale (8-bit, 0-255 range). The lower the number or closer to 0, the darker the stain is. The higher the number or closer to 255, the lighter the stain. The TH intensity was measured on ImageJ, a platform used to analyze digital data.

Results. Figures 2 and 3 presents the visualization of the segmentation results from our best model. As seen, our method can effectively detect and segment the dopaminergic neurons of varying size and shape. Our quantitative results in Table 1, together with the qualitative results in Figures 2 and 3 demonstrate



	Ground Truth (Manual counting) vs. Cellpose
Pearson r	
r	0.9443
95% confidence interval	0.8538 to 0.9794
R squared	0.8917
P value	
P (two-tailed)	<0.0001
P value summary	****
Significant? (alpha = 0.05)	Yes
Number of XY Pairs	18

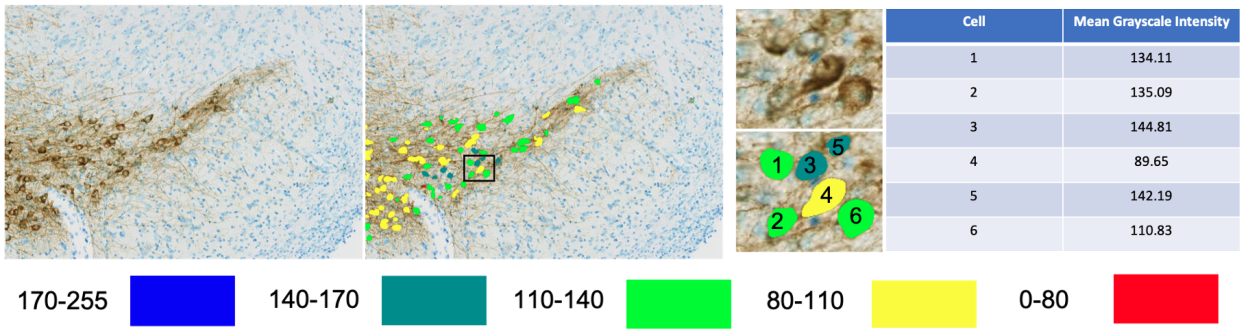
Figure 5: Correlation plot depicting the number of DA neurons counted by a biologist vs the number of DA neurons counted by Cellpose model [9]. Blind data set was used to count the neurons that was previously used to analyze model efficiency in Figure 4.



ANOVA summary	
F	0.1309
P value	0.8776
P value summary	ns
Significant diff. among means (P < 0.05)?	No
R squared	0.005108

Bonferroni's multiple comparisons test	Mean Diff.	95.00% CI of diff.	Below threshold?	Summary	Adjusted P Value
Model counting vs. Ground Truth	-8.556	-67.54 to 50.43	No	ns	>0.9999
Model counting vs. Cellpose	4.278	-54.71 to 63.26	No	ns	>0.9999

Figure 6: Data showing the comparison between Cellpose model and the model developed in this study to count DA neurons in individual sections. The green, black and brown dots depict the cells counted by the model, ground truth (GT) and Cellpose respectively. The red lines indicate the comparison between GT and the Model. The blue lines indicate the comparison between GT and Cellpose. Sections were selected from the blind dataset.



Cell	Mean Grayscale Intensity
1	134.11
2	135.09
3	144.81
4	89.65
5	142.19
6	110.83

Figure 7: Measurement of cell TH intensity. The top panel shows an example image and the color overlay of mean intensity measured in 8-bit grayscale. The bottom panel shows the intensity color legend, a magnified view of several cells outlined in the black box above and a table of mean grayscale intensity value for each cell.

the capability of our framework in providing an effective solution for segmentation of dopaminergic neurons. Figure 4 shows the correlation plot between GT and model counted DA neurons. R^2 of 0.95 with a $pvalue < 0.0001$ was achieved by our model in correlation statistical analysis. Under same parameters and dataset, Cellpose achieved a R^2 of 0.89 with a $pvalue < 0.0001$ in the correlation statistical analysis (see Figure 5). A deeper analysis of the data in Figure 6 revealed that Cellpose undercounted the DA neurons in three sections highlighted by an arrow when compared to the GT counts. Our model was able to count the cells with higher accuracy when compared to GT in these three sections. Figure 7 shows the TH intensity (brown color) of individual DA neuronal cell body in 5 different gradients. The gradient was obtained by measuring the TH intensity for an entire data-set and splitting it into 5 different groups and a visual and numerical data was obtained for each neuron.

5 Conclusion

The goal of this study was to develop a robust machine learning model that can detect and count the DA neurons reliably in independent animal studies. This is an immediate requirement in the field of PD research to accelerate the in-vivo screening of potential drugs so that more drugs can be taken into the clinic for human trials. The existing manual counting or stereology based method is unable to keep up with the number of studies currently conducted in different labs focusing on this area. Additionally, it also suffers from human bias which makes the data interpretation extremely cumbersome. The study framework is established on a self-supervised learning paradigm to combat the lack of large-scale annotated data for training deep models. We also realized that using segmentation based methods facilitated us to go beyond counting the number of DA neurons which the existing machine learning models are implementing. In addition to counting the DA neurons, we were able to obtain characteristics of the DA neurons which is very valuable to understand the health status of individual DA neurons on a histology slide. The loss of dopaminergic neurons determines the extent of pathogenesis in PD. Our model determines the number of dopaminergic neurons by counting the no. of cell body/ soma of dopaminergic neurons. In addition, the ability to measure the TH intensity in individual cells (indicator of dopaminergic neuronal health) helps us to categorize the cells in compartments defining the effect of disease pathology progression in different mouse models. The information we obtain on the health of the neurons is extremely valuable to determine the efficacy of drugs in PD preclinical trials. Studies have shown that TH intensity is lost in specific DA neurons in PD transgenic and injection animal models but the overall analysis of such observation manually is next to impossible [30, 31, 32]. With our method, we can get the cumulative data to dig deeper into the biology of DA neuronal loss and understand the mechanism through which this loss happens and what it suggests in terms of PD pathogenesis. There

are additional challenges to consider such as limited data-sets, tissue section thickness, image resolution and overlapping cells but our model has demonstrated very high efficiency taking into consideration all these factors. With the advancement in machine learning and biology, these models will improve and provide solutions to the ever increasing demand for data-analysis in research biology. Our data suggests that we could extrapolate this method to other species that are used as animal models in PD. With the addition of more dataset, we could go deeper in understanding the biology of DA neuronal loss by capturing the changes which are visible or sometimes not visible to the human eye. To summarize, this method will be very useful to shorten the time needed to analyze loss of DA neurons in animal studies and accelerate the drug discovery of PD.

6 Author Contributions

Fatemeh Haghighi contributed to preparing the data, implementing the method, conducting experiments, analyzing the results, and writing the manuscript. Soumitra Ghosh contributed to supervision, conceptualization, data collection, data analysis, writing and revising the Manuscript. Fatemeh Haghigh and Soumitra Ghosh equally contributed to the study. Hai Ngu contributed to data collection, data analysis, revising the manuscript. Sarah Chu contributed to data preparation and analysis. Han Lin contributed to data preparation, conducting experiments. Mohsen Hejrati contributed to advising the project, revising the manuscript. Baris Bingol contributed to supervision and conceptualization. Somaye Hashemifar contributed to advising and led the project, conception and design of the work, data preparation, writing and revising the manuscript.

References

[1] Tajbakhsh, N. *et al.* Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Medical Image Analysis* **63**, 101693 (2020). URL https://www.sciencedirect.com/science/article/pii/S136184152030058X.

[2] Moshkov, N., Mathe, B., Kertesz-Farkas, A. & et al. Test-time augmentation for deep learning-based cell segmentation on microscopy images. *Scientific Reports* **10**, 5068 (2020).

[3] Greenwald, N. F. *et al.* Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. *Nature biotechnology* **40**, 555–565 (2022).

[4] Johnson, B. N., Charan, R. A. & LaVoie, M. J. Recognizing the cooperative and independent mitochondrial functions of parkin and pink1 (2012).

[5] Ghosh, S. *et al.* α-synuclein aggregates induce c-abl activation and dopaminergic neuronal loss by a feed-forward redox stress mechanism. *Progress in Neurobiology* **202**, 102070 (2021).

[6] Poewe, W. *et al.* Parkinson disease. *Nature Reviews Disease Primers volume* **3**, 17013 (2017).

[7] Guatteo, E., Berretta, N., Monda, V., Ledonne, A. & Mercuri, N. B. Pathophysiological features of nigral dopaminergic neurons in animal models of parkinson disease. *International Journal of Molecular Sciences* **23** (2022). URL https://www.mdpi.com/1422-0067/23/9/4508.

[8] Stringer, C., Michaelos, M. & Pachitariu, M. Cellpose: a generalist algorithm for cellular segmentation. *Nature Methods* **18**, 100–106 (2020).

[9] Robitaille, M. C., Byers, J. M., Christodoulides, J. A. & Raphael, M. P. Robust optical flow algorithm for general single cell segmentation. *Plos one* **17**, e0261763 (2022).

[10] Haghighi, F., Hosseinzadeh Taher, M. R., Gotway, M. B. & Liang, J. Dira: Discriminative, restorative, and adversarial learning for self-supervised medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022).

[11] Azizi, S. *et al.* Big self-supervised models advance medical image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 3478–3488 (2021).

[12] Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241 (Springer, 2015).

[13] Falk, T., Mai, D., Bensch, R., cicek, o. & Abdulkadir, e. a., Ahmed. U-net: deep learning for cell counting, detection, and morphometry. *Nature Methods* **16**, 67–70 (2019).

[14] Hatipoglu, N. & Bilgin, G. deep learning algorithms by utilizing spatial relationships. *Medical & Biological Engineering & Computing volume* **55**, 1829–1848 (2017).

[15] Penttinen, A.-M. *et al.* Implementation of deep neural networks to count dopamine neurons in substantia nigra. *European Journal of Neuroscience* **48**, 2354–2361 (2018). URL https://onlinelibrary.wiley.com/doi/abs/10.1111/ejn.14129. https://onlinelibrary.wiley.com/doi/pdf/10.1111/ejn.14129.

[16] Zhao, S., Wu, K., Gu, C., Pu, X. & Guan, X. Snc neuron detection method based on deep learning for efficacy evaluation of anti-pd drugs. In *2018 Annual American Control Conference (ACC)*, 1981–1986 (2018).

[17] Zbontar, J., Jing, L., Misra, I., LeCun, Y. & Deny, S. Barlow twins: Self-supervised learning via redundancy reduction. *arXiv:2103.03230* (2021).

[18] He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020).

[19] Caron, M. *et al.* Unsupervised learning of visual features by contrasting cluster assignments. *arXiv:2006.09882* (2021).

[20] Grill, J.-B. *et al.* Bootstrap your own latent - a new approach to self-supervised learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F. & Lin, H. (eds.) *Advances in Neural Information Processing Systems*, vol. 33, 21271–21284 (Curran Associates, Inc., 2020). URL https://proceedings.neurips.cc/paper/2020/file/f3ada80d5c4ee70142b17b8192b2958e-Paper.pdf.

[21] Chen, X. & He, K. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15750–15758 (2021).

[22] Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. In III, H. D. & Singh, A. (eds.) *Proceedings of the 37th International Conference on Machine Learning*, vol. 119 of *Proceedings of Machine Learning Research*, 1597–1607 (PMLR, 2020). URL https://proceedings.mlr.press/v119/chen20j.html.

[23] Hosseinzadeh Taher, M. R., Haghighi, F., Feng, R., Gotway, M. B. & Liang, J. A systematic benchmarking analysis of transfer learning for medical image analysis. In *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health*, 3–13 (Springer International Publishing, Cham, 2021).

[24] Haghighi, F., Taher, M. R. H., Zhou, Z., Gotway, M. B. & Liang, J. Transferable visual words: Exploiting the semantics of anatomical patterns for self-supervised learning. *IEEE Transactions on Medical Imaging* **40**, 2857–2868 (2021).

[25] Haghighi, F., Hosseinzadeh Taher, M. R., Zhou, Z., Gotway, M. B. & Liang, J. Learning semantics-enriched representation via self-discovery, self-classification, and self-restoration. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, 137–147 (Springer International Publishing, Cham, 2020).

[26] Chaitanya, K., Erdil, E., Karani, N. & Konukoglu, E. Contrastive learning of global and local features for medical image segmentation with limited annotations. In *Advances in Neural Information Processing Systems*, vol. 33, 12546–12558 (Curran Associates, Inc., 2020). URL https://proceedings.neurips.cc/paper/2020/file/949686ecef4ee20a62d16b4a2d7ccca3-Paper.pdf.

[27] Kaku, A., Upadhya, S. & Razavian, N. Intermediate layers matter in momentum contrastive self supervised learning (2021). 2110.14805.

[28] Li, B., Li, Y. & Eliceiri, K. W. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14318–14328 (2021).

[29] Raghu, M., Zhang, C., Kleinberg, J. & Bengio, S. Transfusion: Understanding transfer learning for medical imaging. In Wallach, H. *et al.* (eds.) *Advances in Neural Information Processing Systems*, vol. 32 (Curran Associates, Inc., 2019). URL https://proceedings.neurips.cc/paper/2019/file/eb1e78328c46506b46a4ac4a1e378b91-Paper.pdf.

[30] Kawahata, I. & Fukunaga, K. Degradation of tyrosine hydroxylase by the ubiquitin-proteasome system in the pathogenesis of parkinson's disease and dopa-responsive dystonia. *International journal of molecular sciences* **21**, 3779 (2020).

[31] Pajarillo, E., Rizor, A., Son, D.-S., Aschner, M. & Lee, E. The transcription factor rest up-regulates tyrosine hydroxylase and antiapoptotic genes and protects dopaminergic neurons against manganese toxicity. *Journal of Biological Chemistry* **295**, 3040–3054 (2020).

[32] Vecchio, L. M. *et al.* Enhanced tyrosine hydroxylase activity induces oxidative stress, causes accumulation of autotoxic catecholamine metabolites, and augments amphetamine effects in vivo. *Journal of neurochemistry* **158**, 960–979 (2021).