Multi-detector ensemble approach for instance segmentation of neuronal cells in microscopic images

Itay Niv
206811309
itayniv10@gmail.com

Sagi Ben Itzhak
307885152
sagibeit@gmail.com

March 16, 2022

Abstract

Neurological disorders are one of the biggest public health challenges of the 21st century, affecting nearly one in every six people worldwide. The prevalence of these conditions is expected to increase significantly as the life expectancy of the world's population continues to grow. Consequently, the need for innovative solutions for the treatment and prevention of these conditions is at an all-time high. One of the most important tasks in drug discovery and development is to accurately quantify the effects of new drugs on their target cells. To that end, cell segmentation techniques are often applied on 2D microscopic images of cell cultures in the analysis process of new drugs. In this work, we proposed a multi-detector ensemble approach for instance segmentation of neuronal cells in microscopic images. Our method employs multiple detection models simultaneously to accurately locate cell instances and uses a transformer-based segmentation module to predict a segmentation mask for each cell. Furthermore, our method combines supervised and semi-supervised approaches to utilize a significant amount of unlabeled data in the training process. With minimal optimization, light-weight backbones, and relatively short training time our approach achieves competitive results in the recent "Sartorius – Cell Instance Segmentation" Kaggle competition achieving 33.2 mAP@0.5:0.95, and ranking in the top 5% in the competition leaderboard. The code and models are available at GitHub repository.

Key words: Instance Segmentation, Neuronal Cells, Cell segmentation, Deep Learning

1 Introduction

Neurological disorders are the leading cause of disability and the second leading cause of death worldwide [1]. Alzheimer's dementia, Parkinson's disease, Multiple Sclerosis, and other neurological disorders affect nearly a sixth of the world's population [2], and the prevalence of many of these disorders drastically increases with age. With the global increase in life expectancy, among other factors, the cost and incidence of neurological disorders is expected to increase significantly as the population aged 60 and over is projected to more than double by 2050[3]. Thus, the demand for innovative solutions and new treatments that can treat or prevent neurological disorders continues to grow. However, quantifying how these conditions respond to treatment is not an easy task. One accepted method is to review neuronal cells via cell microscopy images. Unfortunately, detection and segmentation of individual neuronal cells in microscopy images can be time-consuming and challenging. Therefore, accurate instance segmentation techniques of neuronal cells can assist physicians and researchers to quantify changes in these cells, potentially leading to new and effective drug discoveries for neurological disorders.

Instance segmentation is a computer vision task that performs both object detection and segmentation. It is one of the biggest challenges compared to other segmentation techniques. Aimed at predicting the object class-label and the pixel-specific object instance-mask, it localizes different classes of object instances present in various images. Instance segmentation techniques are essential in many domains, including robotics, medical imaging, autonomous driving, and surveillance. Automating this process is not easy, because the number of instances is not known in advance. With the development of new deep neural network (DNN) techniques, supervised deep learning methods have achieved impressive performance in a wide range of computer vision problems, including object detection ([4, 5, 6]), semantic segmentation ([7, 8, 9]), and instance segmentation ([10, 11, 12]).

In this work, we introduce a multi-detector ensemble approach for instance segmentation of neuronal cells called MDE-CellMask. A novel method that classifies, detects and segments individual neuronal cells in microscopic images. Our approach combines supervised and semi-supervised methods and ensembles multiple detection models to achieve impressive performance with a relatively small number of annotated images.

This paper is organized as follows. In Sec. 2, we review previous work in the field of cell instance segmentation. Sec. 3 provides the details about the two datasets we experimented on: the ”Sartorius” competition dataset and the LIVEcell dataset. In Sec. 4, the general framework of MDE-CellMask is explained, and the methods it is consisted of are reviewed. Details regarding the evaluation protocol, implementation details, experiments, and results are given in Sec. 5. Conclusions and future work are outlined in Sec. 6.

2 Related Work

Segmentation of individual cells and their compartments is a challenging task that is essential for many biological and medical applications. Segmentation of neuronal cells has proven to be especially difficult due to their unique morphology. In the last decade, with the development of DNNs, deep learning-based methods have achieved remarkable progress in object detection and segmentation tasks, leading to advancements in many research areas, including cell segmentation. To name a few, Stringer et al. proposed Cellpose[13], a generalized algorithm for cell segmentation with the ability to segment a wide range of images and cell types. The Cellpose network is a modification of the U-net[14] architecture that utilizes the horizontal and vertical gradient maps to predict the label map of the original image. Yi et al.[15] introduce an attentive instance segmentation model for neural cells based on a single-shot multibox detector (SSD) architecture [16] for the detection task, and U-net for the segmentation task. In their work, they incorporate an attention module in SSD to help it focus on useful image regions while suppressing irrelevant background information. To perform the segmentation, bounding box predictions for the cell instances from the SSD are cropped and fed to a modified U-net.

While the above methods have achieved impressive results, they have a few limitations that are important to address:

- First, both methods rely solely on architectures based on convolutional neural networks (CNNs). Convolutional kernels process a local neighborhood in space. Therefore, while CNNs have dominated many computer vision tasks in the last decade, they are limited in their ability to capture long-range dependencies. The problem of long-range interactions has been tackled in sequential tasks through the use of transformers [17]. In the past few years, transformers have been employed in computer vision models as well, achieving state-of-the-art performance in many computer vision tasks. In our work, we propose a model that integrates the Swin-Transformer[18] in both the detection task and the segmentation task to boost the performance of our model.

- The second limitation is the lack of use of unlabeled data. Acquiring relevant data is an important and challenging stage for medical and biological applications. The performance of machine learning models is greatly dependent on the quality and quantity of the data used for training. Establishing medical and biological datasets is challenging and expensive, as the process requires professionals in the relevant fields, alongside ethical and legal conflicts. Consequently, these datasets tend to be smaller in size compared to other domains. Semi-supervised and unsupervised methods, which make use of unlabeled data in the training process, are widely used to boost performance when data is scarce. In contrast to Cellpose and the method introduced by Yi et al., which use supervised approaches, we combine semi-supervised and supervised approaches using ensemble modeling for the detection task.

3 Data

In this work, we use two different microscopic image datasets of individual cells in 2D cell cultures. The first is the ”Sartorius” competition dataset, and the second is the LIVEcell dataset, which we will expand on in Sec. 3.1 and Sec. 3.2.

3.1 Competition Dataset

The first dataset is introduced as part of the “Sartorius – Cell Instance Segmentation” Kaggle competition [19] which launched on October 14 in 2021 and will be referred to in this work as the “competition dataset”. The competition dataset (Fig. 1) consists of 2818 microscopic images with predefined splits into a training set and test set. It is comprised of three different neuronal cell types, namely SH-SY5Y (neuroblastoma), Cort (neurons), and Astro (astrocytes). The training set contains 2578 images and is comprised of 606 annotated images with a total of 73582 individual cells, and 1972 images without annotations. The test set contains 240 annotated images.

3.2 LIVEcell Dataset

The LIVEcell dataset is introduced in [20]. It is a large, high-quality, manually annotated, and expert-validated dataset of phase-contrast images, consisting of 5,239 annotated microscopic images with a total of 1,686,352 individual cells. LIVEcell dataset (Fig. S1) is comprised of eight different cell types, namely A172 (glioblastoma), BV-2 (microglia), BT-474 (breast cancer), Huh7 (hepatocellular), MCF7 (breast cancer), SkBr3 (Ovarian cancer), SH-SY5Y (neuroblastoma), SK-OV-3 (breast cancer).

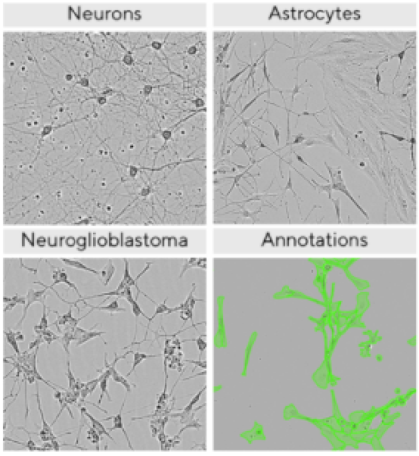

Figure 1: Image examples and annotations from the Competition dataset.

Notice that in both the LIVEcell dataset and the competition dataset, the SH-SY5Y cell class is present.

4 Methods

An overview of MDE-CellMask is depicted in Fig. 2. The proposed architecture can be divided into two main components, namely, the detection module and the segmentation module. The former is comprised of four detection models referred to as “branches”. The first three branches contain supervised models, which include a Faster R-CNN model [4] and two Cascade R-CNN models [21], which will be presented in Sec. 4.1.1 and Sec. 4.1.2, respectively. The last branch contains a semi-supervised soft teacher model [22] that we will expand on in Sec. 4.1.3. The input image is resized to 540x720 before entering the detection module, where it is fed to the four branches simultaneously. Each detection branch generates bounding box predictions for the input image independently of the other branches. The predictions from all four branches are then ensembled using weighted boxes fusion (WBF) [23], which we will present in Sec. 4.1.4. Finally, the ensembled predictions are fed to the segmentation module (Sec. 4.2), where mask predictions are generated for all cell instances in the input image.

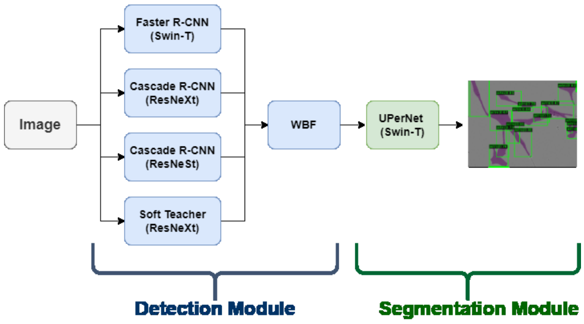

Figure 2: An overview of our multi-detector ensemble approach for instance segmentation of neuronal cells (MDE-CellMask).

4.1 Detection

4.1.1 Faster R-CNN

Faster R-CNN is a single-stage object detection architecture introduced in 2015 [4] as the third iteration of R-CNN papers, followed by the original R-CNN [24] and the Fast R-CNN [25]. Since then, it has been one of the most widely used object detection methods. The Faster R-CNN architecture consists of three main stages. The first stage is a feature extraction network, often referred to as the "backbone" network. The second stage is the region proposal network (RPN), which is used for generating object proposals in the form of bounding boxes. The third stage performs classification and box regression on the proposals from the previous stage.

In this work, for the backbone network of the Faster R-CNN model, we use the Swin-T transformer (Fig 5)[18]. The Swin Transformers are a family of vision transformers introduced by Lui et al. in 2021. It can be used as a general-purpose backbone for computer vision problems, such as image classification and dense recognition tasks. The Faster R-CNN model serves as one of the detection branches in the detection module as shown in Fig. 2.

4.1.2 Cascade R-CNN

Cascade R-CNN is a multi-stage extension of the Faster R-CNN object detection architecture introduced by Zhaowei Cai et al. in 2017 [21]. It seeks to address problems with degrading performance with increased IoU thresholds. The architecture (Fig. 3) consists of a sequence of Faster R-CNN detectors trained with increasing IoU thresholds to be sequentially more selective against close false positives. The detectors are trained stage by stage, leveraging the observation that the output of a detector is a good distribution for training the next higher quality detector.

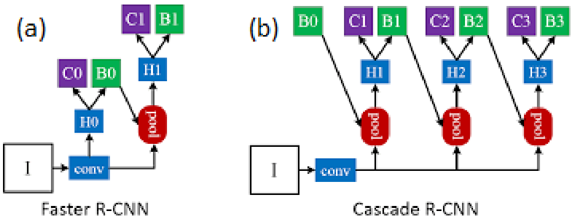

Figure 3: General architecture of (a) Faster R-CNN and (b) Cascade R-CNN frameworks.

In this work, we use the Cascade R-CNN for two models with two different commonly used backbone networks:

- The first backbone network we use is the ResNeXt [26], which integrates the multi-path into ResNet (residual neural network) [27]. ResNeXt is constructed by repeating a building block that aggregates a set of transformations with the same topology (Fig. S3). Compared to a ResNet , it exposes a new dimension, cardinality (the size of the set of transformations), as an essential factor in addition to the dimensions of depth and width.

- The second backbone network we use is the ResNeSt [28], a variant on the ResNeXt that instead stacks Split-Attention blocks. The cardinal group representations are then concatenated along the channel dimension (Fig. S4).

Additionally, we applied the Cascade Mask R-CNN as a baseline in this study, which extends the Cascade R-CNN to instance segmentation by adding a mask head to the cascade. It serves as a baseline against which our MDE-CellMask is compared.

4.1.3 Soft Teacher

The fourth branch of the detection module is a pseudo-label based, semi-supervised, end-to-end framework [22], which follows the teacher-student training scheme. It discards the complicated multi-stage schema adopted by previous pseudo-label approaches [29, 30]. The method simultaneously improves the detector and pseudo labels by leveraging a student model for detection training, and a teacher model that is continuously updated by the student model through the exponential moving average (EMA) strategy for online pseudo-labeling. Within the end-to-end training, two simple techniques are presented, named soft teacher and box jittering to facilitate the efficient leverage of the teacher model.

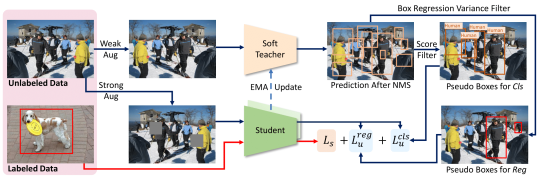

Figure 4: The overview of the end-to-end pseudo-labeling framework for semi-supervised object detection.

In each training iteration, it can be seen in Fig. 4, that labeled images and unlabeled images are randomly sampled (according to a data sampling ratio) to form a training data batch. The teacher model generates the pseudo boxes on unlabeled images, and the student model is trained on both labeled images with the ground-truth and unlabeled images with the pseudo boxes as the ground-truth. Thus, the overall loss is defined as the weighted sum of the supervised loss and the unsupervised loss. Strong augmentation is applied for detection training of the student model, and weak augmentation is used for pseudo-labeling of the teacher model. On the MS-COCO benchmark, this method outperforms the state-of-the-art methods in partially labeled data.

In this work, we use the Faster R-CNN as our detection architecture for the Soft Teacher framework, as done in [22]. For better performance, ResNeXt is adopted as the backbone network.

4.1.4 Weighted Boxes Fusion

Weighted Boxes Fusion (WBF) is a technique used for merging overlapping bounding box predictions. It is commonly used for ensembling predictions in object detection tasks, as it has been shown to achieve superior performance compared to non-max suppression (NMS) and soft-NMS when applied to bounding box predictions originating from multiple detection models [23]. Unlike NMS methods that simply remove part of the predictions, WBF uses information from all bounding boxes to generate the final predictions. In brief, predictions are clustered into groups based on an overlapping criterion defined by a predetermined IoU threshold (hyperparameter). Predictions from the same cluster are then fused to generate a single bounding box prediction, which is a weighted average of the boxes in the group. When applying WBF to predictions from multiple models, the weights for the averaging process in each cluster are determined by two factors, the confidence score of predictions and the model they originate from (hyperparameter). The latter is determined by a model weights vector, where each element corresponds to a single model and defines its' influence on the fused box compared to the other models.

In this work, we use WBF in the detection module to ensemble the predictions from the four detection branches as shown in Fig. 2. The bounding box predictions generated by the WBF are fed to the segmentation module where a mask prediction is generated for each cell instance.

4.2 Segmentation

For the segmentation task, we use a single model approach based on the Swin-T architecture (Fig. 5) introduced in 4.1.1. We utilize UperNet[31] as our base framework, as done by Liu et al.[18]. The segmentation module is trained on the ground truth (GT) bounding boxes and masks. Each input image and its corresponding mask are cropped and resized multiple times based on the GT bounding boxes to a fixed size of 128x128 using ROIAlign before entering the Swin-T based predictor. Each cell instance, corresponding to a single bounding box, is passed through the predictor where a binary (i.e. foreground, background) mask prediction is generated. Finally, the mask predictions are pasted according to the locations of their bounding boxes to form the complete mask prediction for the input image.

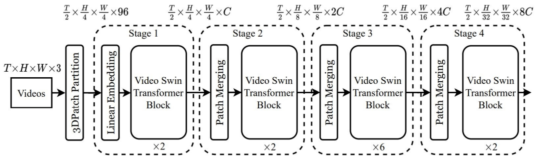
$\frac{T}{2}\times\frac{H}{4}\times\frac{W}{4}\times96$
$\frac{T}{2}\times\frac{H}{4}\times\frac{W}{4}\times C$
$\frac{T}{2}\times\frac{H}{8}\times\frac{W}{8}\times 2C$
$\frac{T}{2}\times\frac{H}{16}\times\frac{W}{16}\times 4C$
$\frac{T}{2}\times\frac{H}{32}\times\frac{W}{32}\times 8C$
Stage 1
Stage 2
Stage 3
Stage 4
$T\times H\times W\times 3$
Videos
3DPatch Partition
Linear Embedding
Video Swin Transformer Block
Patch Merging
×2
×2
×6
×2

Figure 5: Overall architecture of Swin transformer (tiny version, referred to as Swin-T).

5 Experiments

This section first describes the evaluation protocol of the Kaggle competition. Second, the training strategy employed for each model is explained. Then, the implementation details of the training and inference pipelines, as well as the ablation study, are presented. Finally, the results of the different approaches and models are reported and compared to the proposed MDE-CellMask.

5.1 Evaluation Protocol

The Kaggle competition metric [32] is used to evaluate performance. We utilized the mask-level average precision (AP) [33] at IoU (intersection-over-union) threshold of α, following the existing works [34, 11, 10, 35, 36]. The metric is known as $AP@\alpha$. Our Kaggle competition metric is based on the COCO competition evaluation protocol [37], where AP is the mean average precision (p) at 10 distinct IoU thresholds, with values ranging from 0.5 to 0.95 and a step size of 0.05:

$$AP = \frac{1}{10} \sum_{l \in \{0.5, 0.55, \cdots, 0.95\}} \max(p(r \geq l)). \tag{1}$$

We gather the predictions of the proposed model, each of which includes a bounding box, a confidence score, and a segmentation mask, in order to compute the $AP@\alpha$. Then, in decreasing order, we sort the predictions by their confidence scores, ensuring that the ones with the highest scores are taken into account first. The maximum IoU between the predicted segmentation mask and all ground-truth masks is calculated for each prediction:

$$IoU = \max_i (\frac{x \cap y_i}{x \cup y_i}). \tag{2}$$

where $x \in \mathcal{R}^{W\times H}$ is the predicted segmentation mask, $y_i \in \mathcal{R}^{W\times H}$ is the i-th ground-truth mask, W and H are the image width and height. A prediction is considered a true positive instance if the maximum IoU is greater than a threshold α, the corresponding ground-truth is marked as detected. The AP metric summarizes the shape of the precision/recall curve and provides an evaluation that measures the performance for both instance detection and segmentation.

5.2 Training strategy

Our multi-detector ensemble model consists of four detection models (Sec. 4.1) and one segmentation model (Sec. 4.2). All four detection models were trained separately and independently from other models. The segmentation model was trained on the GT bounding boxes. The backbones for all models were initialized with parameters pre-trained on ImageNet. All models were first pretrained on the LIVEcell dataset and then fine-tuned on the competition data.

5.3 Implementation Details

5.3.1 Training

The training data for the LIVEcell pre-train stage consists of the full LIVEcell dataset (train, validation, and test splits). The training and test splits were concatenated to form the train set, while the validation set was used for evaluation. The three supervised detection branches and the Faster R-CNN component in the Soft Teacher branch were pretrained on the LIVEcell train set for 30-40 epochs. The segmentation model's pixel-specific object instance-mask predicting procedure is more computationally expensive to train than the predicting process for the object's bounding box and class label. As a result, the segmentation model was trained on the LIVEcell training set for only 5 epochs.

The training data for the competition consists of 606 annotated images and additional 1972 unannotated images. The annotated portion of the training data was split into a training set and a validation set with a ratio of 95%-5%. The three supervised branches and the segmentation model were fine-tuned for another 30-40 epochs on the training set of the annotated data. The semi-supervised Soft Teacher branch with the pre-trained Faster R-CNN component was trained for another 30-40 epochs on the training set of the annotated portion, and the entire unannotated portion with a sampling ratio (labeled:unlabeled) of 1:1 in each training batch as described in section 4.1.3.

Data augmentations for the supervised detection models included: random flip, shift, scale, rotate, random brightness contrast, gaussian blur, and median blur. Data augmentations for the segmentation model included random flip and rotate. For the semi-supervised Soft Teacher model, the default augmentations from [22] were used. All three supervised detection models and the segmentation model were optimized using AdamW [38] with a mini-batch size of 2, weight decay of 10^{-2}, an initial learning rate of 10^{-4}, and Cosine Annealing as the learning rate policy. The semi-supervised Soft Teacher model was optimized using SGD with a mini-batch size of 2, initial learning rate of 10^{-3}, momentum of 0.9, weight decay of 10^{-4}, and a step learning rate policy with a drop factor of 10 after the 8th and 11th epoch.

5.3.2 Inference

The inference pipeline contains a forward pass of the input image to the four detection branches simultaneously. Test time augmentation (TTA) includes random flip. The predictions of the four detection models are ensembles using WBF as described in section 4.1.4 with a model weights vector of [2,3,3,1] (where the order of the weights corresponds to the order in Fig. 2 from top to bottom), IoU threshold of 0.6, and a skip box threshold of 0.01. The ensembled boxes are used to extract the cell instances from the input image. Each cell instance is cropped and resized as explained in section 4.2, before entering the forward pass of the segmentation model to generate a mask prediction. Finally, mask predictions for all cell instances in the input image are pasted according to the location of their bounding box predictions to form the complete mask prediction for the input image. Post-processing includes instance exclusion by confidence threshold per class with the following thresholds: Shysy5y - 0.25, Cort - 0.55, Astro - 0.35 (i.e. excluding cell instances with a confidence score below the class threshold).

5.4 Ablation Study

For the ablation study, we created four distinct single-detector models using the four detection branches from the detection module. Each single-detector model is comprised of a detection module containing a single detection branch and a segmentation module, as shown in Fig. 6. The single-detector models corresponding to branches 1, 2, and 4 were trained and evaluated with and without the LIVEcell pre-train stage to investigate the effect of the LIVEcell pre-train stage.

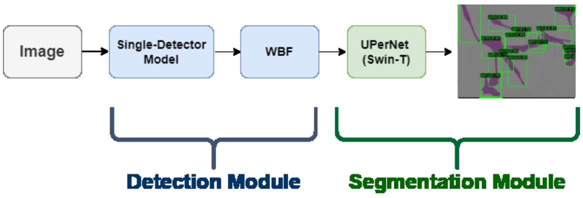

Figure 6: An overview of the framework using a single-detector model.

5.5 Results

We report the results on the competition test set for our MDE-CellMask, the four single-detector models with and without LIVEcell pre-train stage, and the Mask R-CNN baseline in Table. 1. We can observe that MDE-CellMask out-performs all single-detector models and the Mask R-CNN baseline model by a large margin. When comparing the single-detector models to the Mask R-CNN baseline, all single-detector models out-perform the Mask R-CNN baseline as well. In addition, it can be observed that the single-detector models with the LIVEcell pre-train stage surpass their corresponding models without the LIVEcell pre-train stage in performance.

In Fig. S2 detection examples are shown. It can be seen that the same image of the cort cell-type class is presented with the detection results of: Soft Teacher semi-supervised model, the Cascade R-CNN supervised model, and the WBF ensemble of the two detection models.

Segmentation examples of the proposed method are presented in Fig. 7. An image example of each of the three cell-type classes, as well as its instance segmentation, is displayed.

Table 1: Results of neuronal cell instance segmentation on the test set of the Competition data.

Method			**LIVEcell pretrain**	**mAP @0.5:0.95**
Detection		**Segmentation**		
Framework	**Backbone**			
Cascade Mask R-CNN (ResNeXt 101)			-	30.6
Faster RCNN	Swin-T	UperNet Swin-T	-	30.6
			V	31.2
Cascade R-CNN	ResNeSt 101	UperNet Swin-T	-	31.1
			V	31.9
Cascade R-CNN	ResNeXt 101	UperNet Swin-T	V	31.8
Soft Teacher Faster R-CNN	ResNeXt 101	UperNet Swin-T	-	30.9
			V	31.1
Ensemble Model		**UperNet Swin-T**	**V**	**33.2**

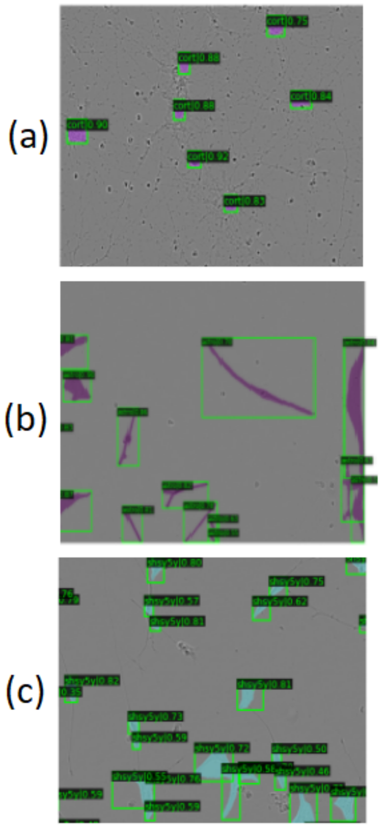

Figure 7: Segmentation examples generated by MDE-CellMask for images of (a) Cort (b) Astro and (c) SH-SY5Y cell types.

6 Conclusions

In this work, we proposed MDE-CellMask, a multi-detector ensemble approach for instance segmentation of neuronal cells. Using supervised and semi-supervised methods, and integrating convolutional-based architectures alongside the Swin transformer, our approach achieves impressive results, outperforming all single-model and single-detector models examined in this paper. With minimal optimization, light-weight backbones, and relatively short training time, our approach achieves competitive results in the recent “Sartorius” Kaggle competition ranking top 5% in the leaderboard. These results demonstrate the great potential of the MDE-CellMask framework in the field of cell segmentation.

For future work, the incorporation of other detection methods as branches in the detection module, as well as the inclusion of a fusion framework for the segmentation task, is an intriguing follow-up research, particularly those based on different methodologies. Secondly, there is a lot of potential in exploring different networks for the semi-supervised, pseudo-labeling framework for training. Finally, we believe that with larger resources and additional optimization, the MDE-CellMask can become a leading framework for cell instance segmentation tasks.

References

[1] A. Avan and V. Hachinski, "Stroke and dementia, leading causes of neurological disability and death, potential for prevention," *Alzheimer's & Dementia*, vol. 17, no. 6, pp. 1072–1076, 2021.

[2] "Nearly 1 in 6 of world's population suffer from neurological disorders – UN report | | UN News," jan 27 2007. [Online; accessed 2022-03-14].

[3] Y. Béjot and K. Yaffe, "Ageing population: A neurological challenge.," *Neuroepidemiology*, vol. 52, no. 1-2, pp. 76–78, 2019.

[4] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[5] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016.

[6] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6154–6162, 2018.

[7] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

[8] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1520–1528, 2015.

[9] D. Yang, Q. Huang, L. Axel, and D. Metaxas, "Multi-component deformable models coupled with 2d-3d u-net for automated probabilistic segmentation of cardiac walls and blood," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 479–483, 2018.

[10] K. He, G. Hkioxari, P. Dollar, *et al.*, "Mask r-cnn. 2980-2988," in *IEEE International Conference on Computer Vision (ICCV). Venice, Italy*, 2018.

[11] Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei, "Fully convolutional instance-aware semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2359–2367, 2017.

[12] M. Bai and R. Urtasun, "Deep watershed transform for instance segmentation," *CoRR*, vol. abs/1611.08303, 2016.

[13] C. Stringer, T. Wang, M. Michaelos, and M. Pachitariu, "Cellpose: a generalist algorithm for cellular segmentation," *Nature methods*, vol. 18, no. 1, pp. 100–106, 2021.

[14] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015.

[15] J. Yi, P. Wu, M. Jiang, Q. Huang, D. J. Hoeppner, and D. N. Metaxas, "Attentive neural cell instance segmentation," *Medical image analysis*, vol. 55, pp. 228–240, 2019.

[16] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*, pp. 21–37, Springer, 2016.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[18] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.

[19] "Sartorius - Cell Instance Segmentation | Kaggle." [Online; accessed 2022-03-12].

[20] C. Edlund, T. R. Jackson, N. Khalid, N. Bevan, T. Dale, A. Dengel, S. Ahmed, J. Trygg, and R. Sjögren, "Livecell—a large-scale dataset for label-free live cell segmentation," *Nature methods*, vol. 18, no. 9, pp. 1038–1045, 2021.

[21] Z. Cai and N. Vasconcelos, "Cascade R-CNN: delving into high quality object detection," *CoRR*, vol. abs/1712.00726, 2017.

[22] M. Xu, Z. Zhang, H. Hu, J. Wang, L. Wang, F. Wei, X. Bai, and Z. Liu, "End-to-end semi-supervised object detection with soft teacher," *CoRR*, vol. abs/2106.09018, 2021.

[23] R. Solovyev, W. Wang, and T. Gabruseva, "Weighted boxes fusion: Ensembling boxes from different object detection models," *Image and Vision Computing*, vol. 107, p. 104117, 2021.

[24] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587, 2014.

[25] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015.

[26] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5987–5995, 2017.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[28] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, Z. Zhang, H. Lin, Y. Sun, T. He, J. Mueller, R. Manmatha, M. Li, and A. J. Smola, "Resnest: Split-attention networks," *CoRR*, vol. abs/2004.08955, 2020.

[29] K. Sohn, Z. Zhang, C. Li, H. Zhang, C. Lee, and T. Pfister, "A simple semi-supervised learning framework for object detection," *CoRR*, vol. abs/2005.04757, 2020.

[30] B. Zoph, G. Ghiasi, T.-Y. Lin, Y. Cui, H. Liu, E. D. Cubuk, and Q. Le, "Rethinking pre-training and self-training," *Advances in neural information processing systems*, vol. 33, pp. 3833–3845, 2020.

[31] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 418–434, 2018.

[32] "Sartorius - Cell Instance Segmentation | Kaggle Evaluatin." [Online; accessed 2022-03-12].

[33] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[34] J. Dai, K. He, and J. Sun, "Instance-aware semantic segmentation via multi-task network cascades," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3150–3158, 2016.

[35] J. Yi, P. Wu, D. J. Hoeppner, and D. Metaxas, "Pixel-wise neural cell instance segmentation," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 373–377, 2018.

[36] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[37] "Coco - Common Objects in Context Detection Evaluation." [Online; accessed 2022-03-12].

[38] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

Supplementary Materials

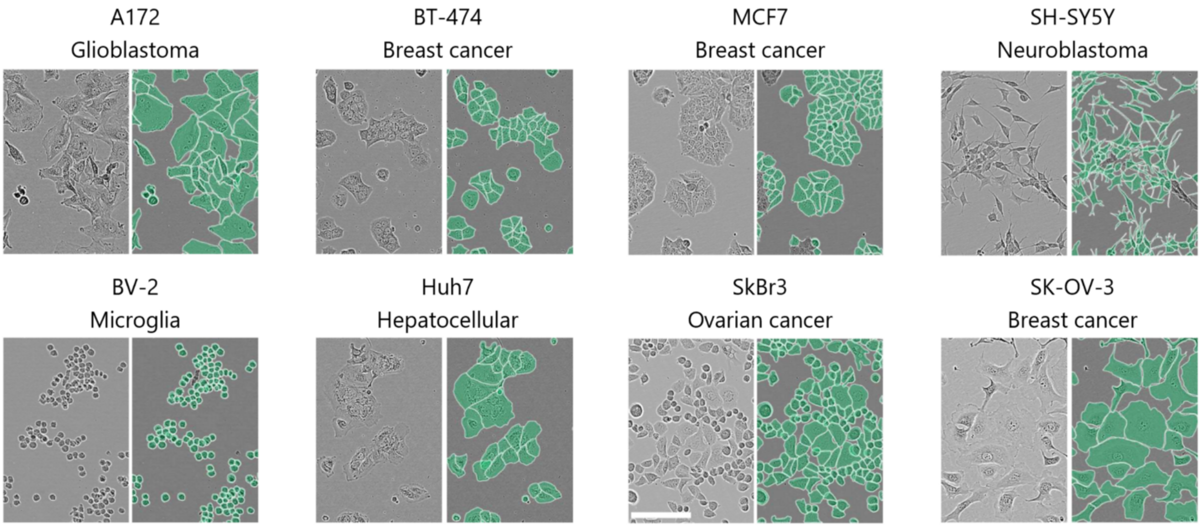

Figure S1: Image examples and annotations from the LIVEcell dataset.

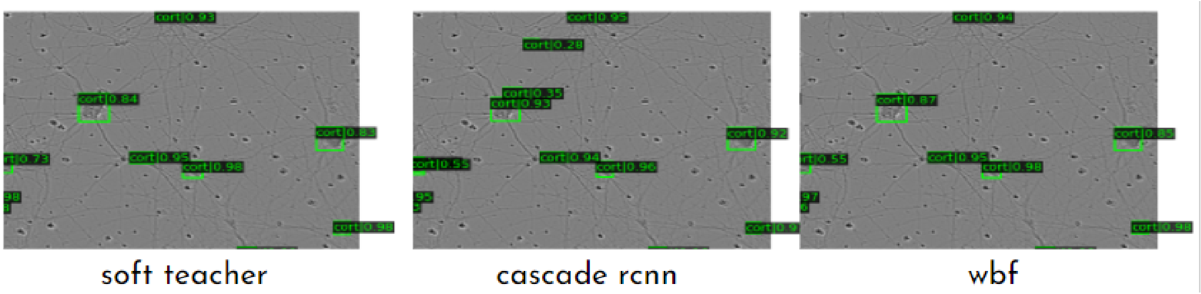

Figure S2: Detection Examples. Left: Soft Teacher semi-supervised model. Middle: Cascade R-CNN supervised model. Left: The WBF ensemble of the two detection models.

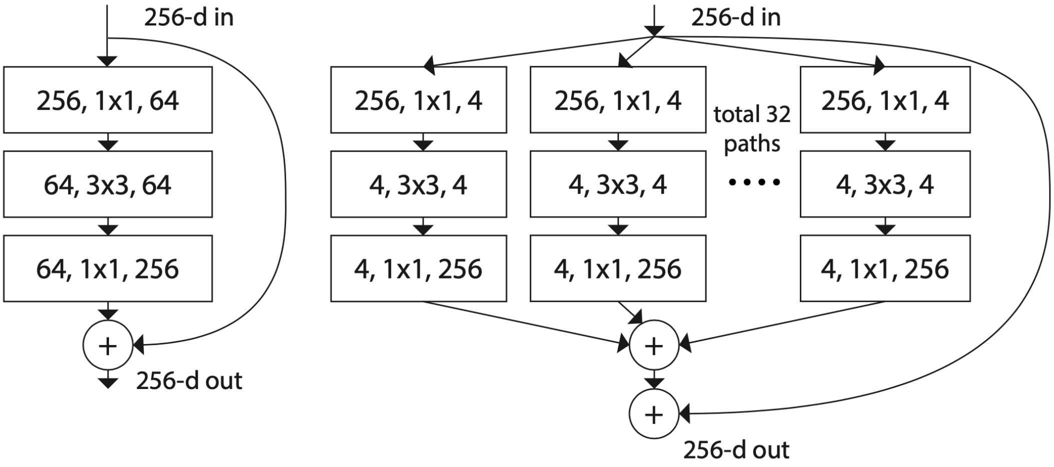

Figure S3: Left: A block of ResNet. Right: A block of ResNeXt with cardinality=32.

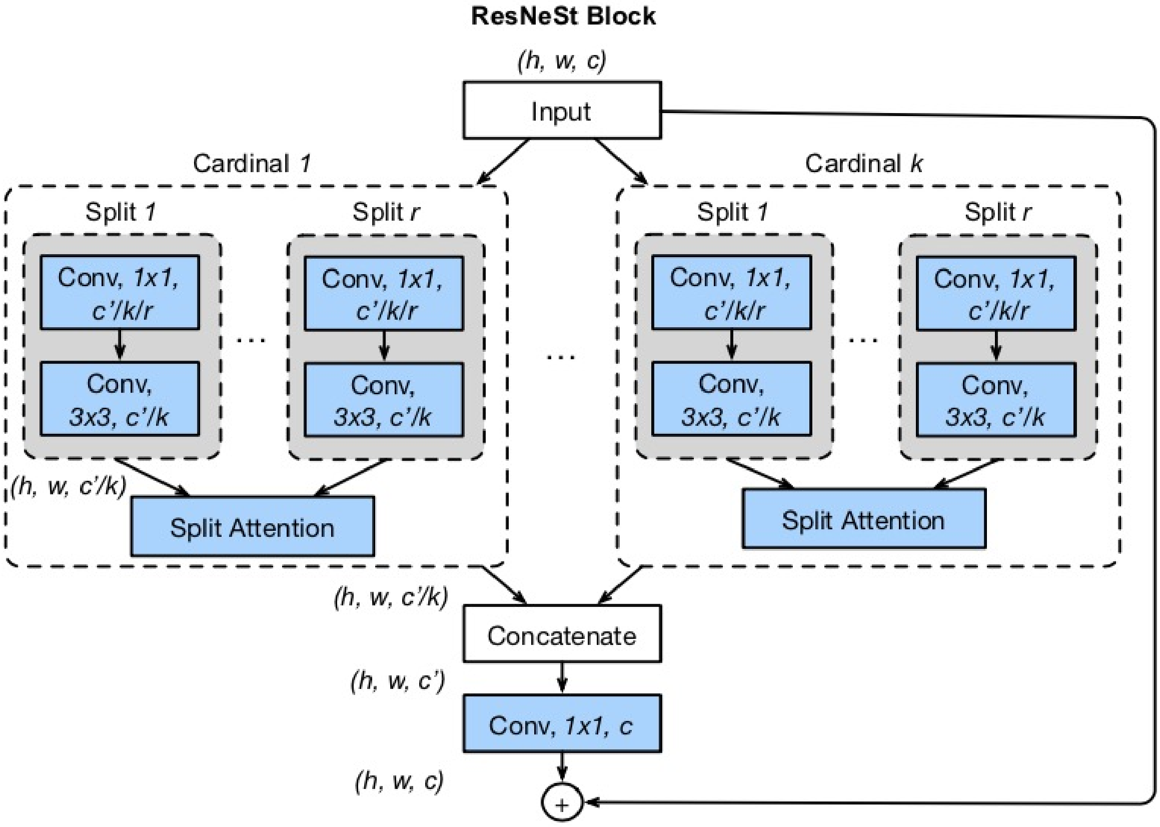

Figure S4: ResNeSt block.