Instance Segmentation of Neural Cells

Jingru Yi[1] [0000−0001−9648−9389], Pengxiang Wu[1] [0000−0002−6929−5877], Menglin Jiang[1] [0000−0002−1959−6161], Daniel J. Hoeppner[2], and Dimitris N. Metaxas[1]

[1] Department of Computer Science, Rutgers University, Piscataway, NJ 08854, USA
{jy486,pw241,menglin.jiang,dnm}@cs.rutgers.edu
[2] Astellas Research Institute of America, San Diego, CA 92121, USA
daniel.hoeppner@astellas.com

Abstract. Instance segmentation of neural cells plays an important role in brain study. However, this task is challenging due to the special shapes and behaviors of neural cells. Existing methods are not precise enough to capture their tiny structures, e.g., filopodia and lamellipodia, which are critical to the understanding of cell interaction and behavior. To this end, we propose a novel deep multi-task learning model to jointly detect and segment neural cells instance-wise. Our method is built upon SSD, with ResNet101 as the backbone to achieve both high detection accuracy and fast speed. Furthermore, unlike existing works which tend to produce wavy and inaccurate boundaries, we embed a deconvolution module into SSD to better capture details. Experiments on a dataset of neural cell microscopic images show that our method is able to achieve better performance in terms of accuracy and efficiency, comparing favorably with current state-of-the-art methods.

Keywords: Neural cell · Instance segmentation · Cell detection · Cell segmentation

1 Introduction

The cellular mechanism involved in the lineage path from a single neural stem cell remains mysterious in neural science. With the aid of real-time microscopy imaging system [15], the specification of neurons, astrocytes, and oligodendrocytes from a single neural stem cell could be recorded as a time-lapse video. As an important tool to explore the interactions between the cells, neural cell instance segmentation algorithm is in great desire since it locates and segments the cells at the same time. In particular, a fast and accurate instance segmentation tool is crucial when we analyze large video datasets. However, neural cell instance segmentation is a challenging problem due to various factors, such as cell mitosis, cell distortion, cell adhesion, unclear cell contours and background impurities. Besides, the tiny and slender structures such as filopodia and lamellipodia involved in cell movement render the problem even more difficult.

Recent years have witnessed a significant improvement in object detection and segmentation due to deep neural network (DNN) techniques [9,10,14,19,21,

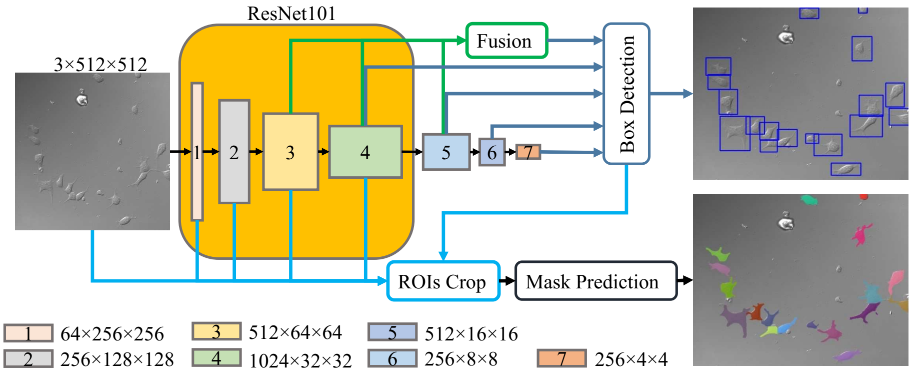

Fig. 1. Overview of our approach. The input image, which has the size of 640 × 512, is resized to 512 × 512 before being fed into the network. The feature maps are displayed as "number of channels × height × width". Block 1-4 are from Residual-101 [8], block 5-7 are the original convolutional blocks of SSD [13].

22]. For example, region-based convolutional network (R-CNN) [5,6,18] was proposed to achieve accurate object detection and classification. To accelerate object detection, the one-stage detector YOLO [16], YOLO9000 [17], and SSD [13] were also proposed. These methods substantially outperform traditional methods [20] which are based on hand-crafted features and classifiers. In the semantic segmentation field, Long *et al.* [14] introduced a ground-breaking fully convolutional network (FCN) that achieves end-to-end, pixel-wise semantic segmentation. Ronneberger *et al.* [19] further extended FCN and proposed a U-Net architecture where successive deconvolutional layers with skip-connections are employed to produce more precise output. To combine both detection and segmentation, i.e., perform instance segmentation, Dai *et al.* [1] proposed a multi-task network cascades (MNC) model that predicts the object box, class, and mask simultaneously. As MNC is time-consuming in prediction, Li *et al.* [11] proposed fully convolutional instance-aware semantic segmentation (FCIS), which predicts the segmentation mask directly from a score map. He *et al.* [7] presented Mask R-CNN, which adds a mask prediction branch to FPN network [12]. However, these methods do not exploit the global context information, which has been proven to be very useful in visual classification tasks [14,19]. Consequently, they fail to accurately predict the fine details of neural cells, such as the filopodia and lamellipodia. Moreover, many of these methods suffer from slow prediction speed. Therefore, they are not suitable for analyzing large microscopic videos.

To overcome the above drawbacks, we propose a novel deep multi-task learning model for neural cell instance segmentation, which takes full advantage of global context information in both detection and segmentation. The overview of our approach is shown in Fig. 1. In particular, our model is based on SSD network [13]. Unlike original SSD, we employ ResNet101 [8] as the backbone instead of VGG network to increase the detection accuracy and speed. To further improve the detection accuracy for fine structures, we utilize a fusion strategy to propagate the context information from the high-level feature maps to the low-level ones. Thanks to the ability of our model to learn the global semantic context, our mask prediction is more precise than the state-of-the-art methods.

2 Methods

The framework of our neural cell instance segmentation approach is illustrated in Fig. 1. The input image is resized to 512×512 before being fed into the network. Note that the predicted boxes range from 0 to 1, and thus the shrinkage of the image does not affect the predictions. Our network jointly predicts the detection bounding box and the segmentation mask for each cell in the image. Below, we first introduce our cell detection module, and then present our cell segmentation module.

2.1 Neural Cell Detection

Our cell detection method builds upon SSD [4, 13]. Unlike original SSD, we replace VGG [4,8] in SSD with ResNet101 network [8] to improve its cell detection accuracy, as ResNet101 is proved to have higher accuracy than VGG network [8]. Moreover, our experiments show that ResNet101-based SSD (0.1017s) runs faster than VGG16-based SSD (0.1537s). The network architecture is shown in Fig. 1. In order to detect cells of different sizes, our box detection module concatenates multi-scale feature maps, which are denoted by blocks 3-7 in Fig. 1. Each feature map is divided into a series of grids, and each grid has the size of 1×1. A grid works as an anchor box that centers in the grid and has a specific scale (i.e., width and height) and aspect ratio. These grids are referred to as default boxes in SSD [13]. As a shallow feature map has a smaller reception field than a deep feature map, the scale of a default box on a shallow feature map is smaller than that on a deep feature map. For example, the scale of a default box on a block 3 feature map is below 0.1, whereas the scale on a block 7 feature map could be as large as 0.75. Finally, following SSD [13], our cell detection module predicts the offsets between the default boxes and the cell bounding boxes with a 3×3 convolutional layer, and predicts the confidence score for each box with another 3×3 convolutional layer.

One drawback of SSD is that its shallow layers contain less semantic information than the deep layers. Consequently, although SSD predicts object locations using multi-scale feature maps, the shallow feature maps could not help detect small objects correctly. To solve this issue and improve our detection accuracy for small cells, we fuse the feature maps in blocks 3-5 and replace the original feature maps in block 3, so as to inject more semantic information to the shallow feature map (see Fig. 1). Specifically, we first use a single 1×1 convolutional layer to transform the feature maps from blocks 3-5 to have the same channel number 256. Then the transformed feature maps from blocks 4-5 are up-sampled to have the same size as the one from block 3 by bilinear interpolation. Finally,

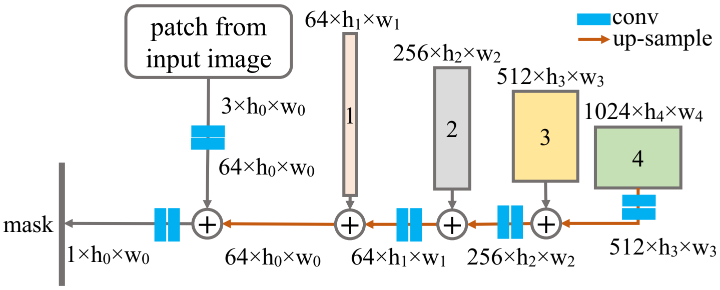

Fig. 2. Architecture of our mask prediction module. The feature maps are displayed as "number of channels × height × width". The convolutional layers are 3×3 with stride 1. Up-sample is bilinear interpolation.

the three transformed feature maps are concatenated together and expanded to have channel number 512 by a 1×1 convolutional layer.

The objective loss for cell detection is a weighted combination of localization loss and confidence loss:

$$L_{\text{det}} = \frac{1}{N_{\text{pos}}}(L_{\text{locs}} + \alpha L_{\text{conf}}), \tag{1}$$

where α is a weight factor, N_{pos} is the number of positive predicted boxes, L_{locs} is a smooth L_1 loss [6] of bounding-box regression offsets [5, 13]:

$$L_{\text{locs}} = \sum_{i \in \text{pos}} \sum_{m \in \{cx,cy,w,h\}} \text{smooth}_{L_1}(l_i^m - g_i^m), \tag{2}$$

where $i \in$ pos denotes the set of positive predicted boxes, and l_i^m and g_i^m refer to the predicted and ground-truth offset boxes, respectively. $m \in \{cx, cy, w, h\}$ indicates the specific localization feature, such as center of the box (cx, cy), width of the box w, and height of the box h. L_{conf} is a binary cross entropy loss between the ground-truth confidence and the predicted box confidence:

$$L_{\text{conf}} = -\sum_i (x_i \log p_i + (1 - x_i) \log(1 - p_i)), \tag{3}$$

where x_i is the ground-truth confidence, and p_i is the predicted box confidence. Particularly, the ground-truth confidence of a default box will be set to 1 if the Jaccard index between this default box and the ground-truth box is greater than 0.5, otherwise the confidence will be set to 0.

2.2 Neural Cell Segmentation

As shown in Fig. 1, after obtaining the bounding box of a cell, we crop the cell box from the input image and feature maps in blocks 1-4, and pass them to our mask prediction module. The architecture of our mask prediction module is shown in Fig. 2. Motivated by FCN [14] and U-Net [19], we combine the shallow layers with deep layers using a single addition operation. In this way, we propagate the context information from deep layers to shallow layers. To make sure two feature maps have the same size when applying the summation operation, we use bilinear interpolation to upsample the crops from deep layers. As the crops are tiny, we also utilize the patch from the input image to take advantage of its finer details. In this way, the details of the crops are reserved, which improves segmentation accuracy. The objective loss of our mask prediction module is a binary-cross entropy loss:

$$L_{\text{masks}} = -\frac{1}{N} \sum_{j}^{N} \sum_{i} (t_{ij} \log p_{ij} + (1 - t_{ij}) \log(1 - p_{ij})), \tag{4}$$

where p_{ij} and t_{ij} are the predicted and ground-truth mask values at position i for the j-th positive predicted bounding box (whose overlap with the ground-truth box exceeds a certain threshold), respectively, and N is the total number of positive predicted bounding boxes.

3 Experiments

3.1 Experimental Settings

Our neural cell image dataset builds on a collection of time-lapse microscopic videos [15]. In particular, we sample 386 images from the videos for training, 129 for validation, and 129 for testing. The image size is 640×512. The ground-truth is labeled by experts. Our method is implemented with PyTorch. During the training process, the ResNet101 network is fine-tuned with the weights pre-trained on ImageNet [2], while other parts of the network are initialized with random weights sampled from a standard Gaussian distribution. To avoid overfitting, we employ data augmentation and early-stop strategy in training. To accelerate the training process, we first train the cell detector. Then we fix the weights of the detection network and train the segmentation network. Note that our model could also be trained in an end-to-end manner. We compare our method with the state-of-the-art instance segmentation algorithms, namely MNC [1], FCIS [11] and Mask R-CNN [7]. All the methods are tested on NVIDIA K40 GPUs.

Following conventions in existing works [1,11], we evaluate the instance segmentation accuracy using average precision (AP) [3] at intersection-over-union (IoU) thresholds of 0.5 and 0.7. In particular, we consider a cell instance segmentation result as a combination of a detection bounding box, a confidence score of the box, and a segmentation mask. During evaluation, all the bounding boxes are sorted by their confidence scores to make sure that boxes with high confidence scores are considered first. For each box, the IoU between its predicted mask and the ground-truth mask is calculated. The box will be considered as a

Table 1. Evaluation results of neural cell instance segmentation. Time is evaluated on a single NVIDIA K40 GPU.

Method	AP@0.5	AP@0.7	IoU@0.5	IoU@0.7	Time (sec)
MNC [1]	48.72	11.37	62.73	75.47	0.4750
FCIS [11]	66.02	7.13	64.85	75.07	0.2130
Mask R-CNN [7]	59.94	25.87	72.10	79.30	0.7486
Ours	**87.39**	**58.38**	**76.23**	**79.64**	**0.1920**

true positive if the IoU score is greater than a threshold (e.g., 0.5 or 0.7), and the corresponding cell is recorded as detected. On the contrary, any repetitive detection or its corresponding mask whose IoU is smaller than the threshold is considered as a false positive. Finally, the AP metric [3] summarizes the shape of the precision/recall curve and measures both instance detection and segmentation accuracy. In addition to AP at mask-IoU, we also measure the average mask IoU at thresholds of 0.5 and 0.7. The computational efficiency of all the methods is also measured according to their testing time.

3.2 Neural Cell Instance Segmentation Results

The evaluation results are summarized in Table 1, which indicates our model outperforms the state-of-the-art methods by a large margin. Several instance segmentation results are provided in Fig. 3 for qualitative evaluation. It can be observed from Fig. 3 that MNC and FCIS are not able to capture the slender and tiny filopodia and lamellipodia of cells. The mask boundaries predicted from FCIS are wavy. Moreover, for images that contain multiple small cells (e.g. the last row in Fig. 3), MNC could not distinguish the cells which are attached or very close to each other, and FCIS is weak in detecting these small cells. The coarse mask prediction and poor detection of smaller cells from MNC and FCIS explain their low AP at mask-IoU of 0.7 (see Table 1). Mask R-CNN is better at capturing tiny structures. However, it fails to capture the long and slender structures. Compared with the state-of-the-art methods, our model learns global semantic context information in both detection and segmentation, thereby exhibiting better performance in detecting small cells and capturing the tiny and slender structures of cells.

4 Conclusion

In this paper, we propose a novel method for neural cell instance segmentation. Compared with existing methods, our model could better detect small cells and capture their tiny and slender structures such as filopodia and lamellipodia. These properties indicate a great potential of our method in neural science research.

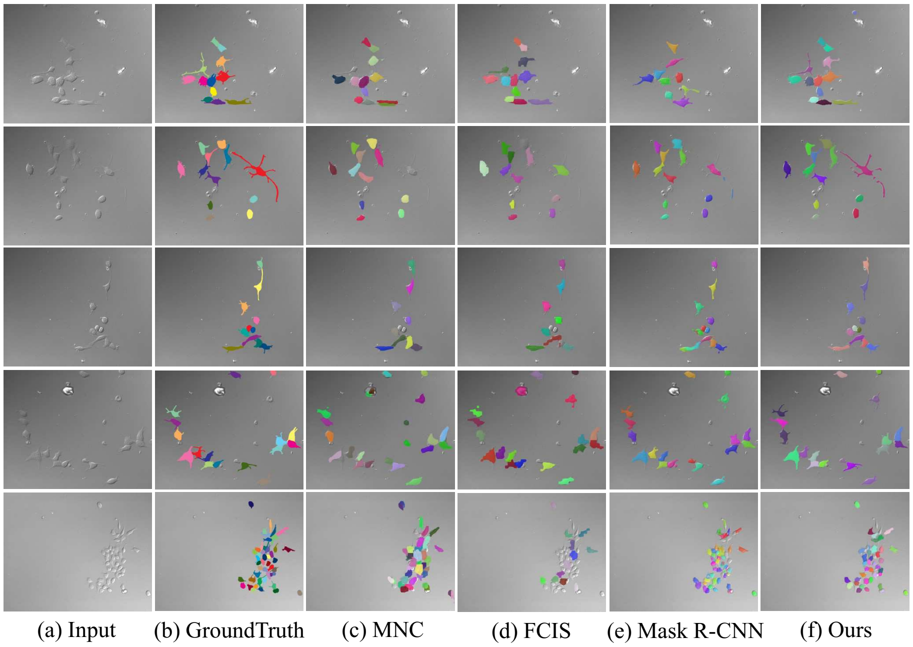

Fig. 3. Neural cell instance segmentation results of MNC [1], FCIS [11], Mask R-CNN [7] and our method. Compared to MNC, FCIS and Mask R-CNN, our method is more accurate and could capture the tiny and slender structures of neural cells, such as filopodia and lamellipodia.

References

1. Dai, J., He, K., Sun, J.: Instance-aware semantic segmentation via multi-task network cascades. In: Proc. IEEE CVPR. pp. 3150–3158 (2016)
2. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Li, F.F.: ImageNet: A large-scale hierarchical image database. In: Proc. IEEE CVPR. pp. 248–255 (2009)
3. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL visual object classes (VOC) challenge. IJCV **88**(2), 303–338 (2010)
4. Fu, C.Y., Liu, W., Ranga, A., Tyagi, A., Berg, A.C.: DSSD: Deconvolutional single shot detector. arXiv:1701.06659 (2017)
5. Girshick, R.B.: Fast R-CNN. In: Proc. IEEE ICCV. pp. 1440–1448 (2015)
6. Girshick, R.B., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proc. IEEE CVPR. pp. 580–587 (2014)
7. He, K., Gkioxari, G., Dollár, P., Girshick, R.B.: Mask R-CNN. In: Proc. IEEE ICCV. pp. 2980–2988 (2017)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proc. IEEE CVPR. pp. 770–778 (2016)
9. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Proc. NIPS. pp. 1097–1105 (2012)

10. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
11. Li, Y., Qi, H., Dai, J., Ji, X., Wei, Y.: Fully convolutional instance-aware semantic segmentation. In: Proc. IEEE CVPR. pp. 4438–4446 (2017)
12. Lin, T.Y., Dollár, P., Girshick, R.B., He, K., Hariharan, B., Belongie, S.J.: Feature pyramid networks for object detection. In: Proc. IEEE CVPR. pp. 936–944 (2017)
13. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C., Berg, A.C.: SSD: Single shot multibox detector. In: Proc. ECCV (1). pp. 21–37 (2016)
14. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proc. IEEE CVPR. pp. 3431–3440 (2015)
15. Ravin, R., Hoeppner, D.J., Munno, D.M., Carmel, L., Sullivan, J., Levitt, D.L., Miller, J.L., Athaide, C., Panchision, D.M., McKay, R.D.G.: Potency and fate specification in CNS stem cell populations in vitro. Cell Stem Cell **3**(6), 670–680 (2008)
16. Redmon, J., Divvala, S., Girshick, R.B., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proc. IEEE CVPR. pp. 779–788 (2016)
17. Redmon, J., Farhadi, A.: YOLO9000: Better, faster, stronger. In: Proc. IEEE CVPR. pp. 6517–6525 (2017)
18. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: Proc. NIPS. pp. 91–99 (2015)
19. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Proc. MICCAI (3). pp. 234–241 (2015)
20. Wu, P., Yi, J., Zhao, G., Huang, Z., Qiu, B., Gao, D.: Active contour-based cell segmentation during freezing and its application in cryopreservation. IEEE TBME **62**(1), 284–295 (2015)
21. Yi, J., Wu, P., Hoeppner, D.J., Metaxas, D.N.: Fast neural cell detection using light-weight SSD neural network. In: Proc. IEEE CVPR Workshop. pp. 860–864 (2017)
22. Yi, J., Wu, P., Hoeppner, D.J., Metaxas, D.N.: Pixel-wise neural cell instance segmentation. In: Proc. IEEE ISBI. pp. 373–377 (2018)