# Instance Segmentation from Volumetric Biomedical Images Without Voxel-Wise Labeling

Meng Dong, Dong Liu(✉), Zhiwei Xiong, Xuejin Chen, Yueyi Zhang, Zheng-Jun Zha, Guoqiang Bi, and Feng Wu

University of Science and Technology of China, Hefei, China
dongeliu@ustc.edu.cn

**Abstract.** Volumetric instance segmentation plays a significant role in biomedical morphological analyses. The improvement of segmentation accuracy has been accelerated by the progress of deep learning-based methods. However, such methods usually rely heavily on plenty of precise annotation, which is time-consuming and may need some expert knowledge to label manually. Although there are several studies focusing on weakly supervised methods in order to save the labeling cost, previous approaches still more or less require voxel-wise annotation. In this paper, we propose a weakly supervised instance segmentation method that needs no voxel-wise labeling. Our approach takes advantage of two advanced techniques: one is the popular proposal-based framework (Faster R-CNN in this paper) for instance detection, and the other is the peak response mapping (PRM) for finding visual cues of instances. Then a new thresholding method combines detected boxes and visual cues to generate final instance segmentation results. We conduct experiments on two biomedical datasets, one of which is a large-scale mouse brain dataset at single-neuron resolution collected by ourselves. Results on both datasets validate the effectiveness of our proposed method.

**Keywords:** Biomedical image analysis · Peak response mapping · Volumetric instance segmentation · Weak supervision

## 1 Introduction

Instance segmentation is a pixel-level visual analysis task, which seeks to not only label precise class-aware masks but also produce instance-aware tags to distinguish same-class individual regions. With accurately segmented instances (e.g. somas), the morphological analyses of biomedical images can be made meticulous and more informative. With the progress of exploring deep learning-based methods for computer vision tasks, the popular multi-task approach [4] achieves excellent performance for instance segmentation on natural images, which performs object detection first and then generates instance masks by the following


D. Shen et al. (Eds.): MICCAI 2019, LNCS 11765, pp. 83–91, 2019.
https://doi.org/10.1007/978-3-030-32245-8_10

mask branch. The approach has been extended and its superiority has been verified for biomedical images [13]. However, exploiting the advanced deep learning methods on biomedical images still faces challenges. One major problem is that these methods usually rely heavily on pixel/voxel-wise detailed labeling, which is laborious and time-consuming especially for volumetric images. Labeling biomedical images may also need some expert knowledge, leading to even higher cost.

Many attempts have been made on biomedical images [1,11,13] aiming at saving labeling cost with weakly- or semi-supervised learning methods. Yang et al. [11] present an active learning method for 2D biomedical image segmentation, which can improve segmentation accuracy through suggesting the most effective rather than all samples for labeling. In [1], a sparse annotation approach is proposed for semantic segmentation from volumetric images: only several slices have pixel-wise labeling because of the structural similarity between sequential 2D images. Zhao et al. [13] apply a modified Mask R-CNN [4] to volumetric data for instance segmentation, and they use bounding boxes for all instances and voxel-wise labels for a small proportion of instances. However, the above mentioned works still more or less demand pixel-wise or voxel-wise annotation.

In fact, there are several existing studies about instance segmentation for natural images without pixel-wise labeling, i.e. with only bounding boxes or image-level classes. A commonly used strategy is self-training: the model is trained in full supervision using labels generated by the model itself in an iterative manner [5], and the rough labels can be refined after several iterations. But these methods are usually sensitive to the initial approximate labels and the iterative procedure is a heavy computation burden. In [9], a visualization method for deep image classification CNN has been explored, which is a top-down attention way. The saliency maps can be extracted by a single back-propagation, and the maps are also used as visual cues for weakly supervised semantic segmentation. Similarly, Zhou et al. [14] come up with a new idea for instance segmentation with only image-level class tags. They use locally class-aware peak response mapping (PRM) results as instance representations, and combine them with a segment proposal retrieving operation to produce the instance segmentation results.

Inspired by the above works, in this paper, we address the problem of 3D instance segmentation from volumetric biomedical images with only bounding-box labeling. We split the task into detection and segmentation, detection can be fulfilled by a deep network-based detector, while segmentation utilizes visual cues from PRM results. But the PRM results are usually not complete to produce segmentation masks, hence we design an advanced thresholding method to employ PRM for segmentation. Our main contributions are as follows:

- To our best knowledge, we propose the first weakly supervised instance segmentation method for volumetric biomedical images that does not rely on any voxel-wise annotation. Instead, our model can be trained with only bounding box annotation.
- We extend the peak response mapping into detection network so as to generate high-quality visual cues to benefit the following thresholding phase.

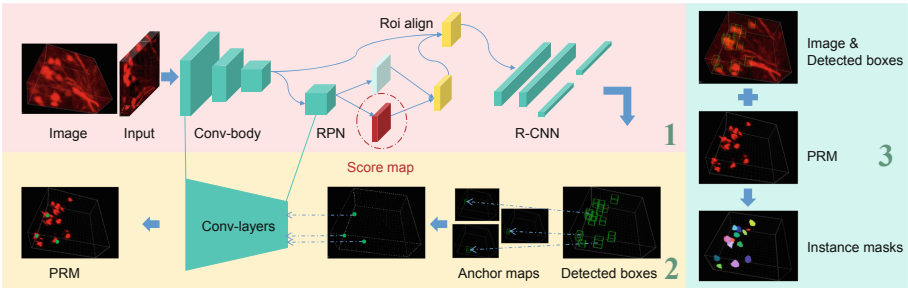


**Fig. 1.** The pipeline of our approach: Part 1 is the detection phase; Part 2 denotes the peak response mapping for extracting visual cues, which is fulfilled by the back-propagation of the anchor locations of the detected boxes through the Conv-layers; Part 3 shows the local thresholding phase. Best viewed in color.

- We design an advanced thresholding method that employs the visual cues extracted from deep learning-based model. And experiments verify that our thresholding method achieves more precise segmentation.
- We provide a mouse brain image dataset at single-neuron resolution, which is acquired by florescence staining and confocal microscopy imaging techniques. We label soma with bounding box for training set, but give voxel-wise mask annotation for testing set.

Our data and code have been published at https://braindata.bitahub.com/.

## 2 Method

The pipeline of our approach is shown in Fig. 1, the primary component is a proposal-based detector (denoted in pink background), which can be trained end-to-end with bounding boxes. Instance segmentation includes three steps: (1) the instances (e.g. cells) are detected as boxes; (2) the visual cues for each instance are obtained by PRM, i.e. the back-propagation from score-map layer to input layer; (3) the final instance masks will be segmented by a local thresholding method which utilizes the detected boxes, visual cues, and image intensity.

***3D Faster R-CNN for Instance Detection.*** We extend Faster R-CNN [8] into 3D version for volumetric image task, including conv-body, region proposal networks (RPN) and region convolutional network (R-CNN). And the roi-align layer is also changed to 3D using trilinear interpolation among 8 neighbor voxels for aligning feature maps for each proposal box. Particularly, for the consideration of reducing computing burden for volumetric data, we select a small but still efficient network as the conv-body of detector for feature extraction, whose structure inherits the down-stream part of DSN [3]. And we may modify this network to adapt to different data. Since Faster R-CNN is a proposal-based

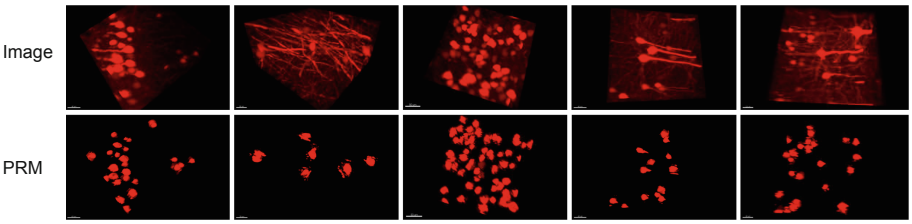


**Fig. 2.** Mouse brain soma data examples and corresponding PRM results.

detector that depends on the default anchors, it is essential to carefully set sizes and aspect ratios for anchors to fit the size distribution of targets. Another important related factor is feature stride, which determines the granularity of sliding anchors. We decrease stride for smaller targets by adjusting conv-body, specifically, removing the last pooling layer and the following layers to reduce stride from 8 to 4. In this way, we can improve the cover rate of anchors and keep more detailed features for small targets. Note that we do not use more complex network structure for the sake of computation burden.

***Peak Response Mapping for Visual Cues.*** PRM implemented in [14] depends on a class peak stimulation layer, which is learnt to predict the class probability corresponding to spatial locations. Inspired by this, we observe that due to the characteristics of detector structure, the score map of RPN is a typical class response map related to locations. Therefore, we assume that those high scores can also indicate the strongly informative voxels, and we propose a new PRM way based on object detection framework. RPN predicts the *score* and *location regression value* for each local anchor, which are used to produce proposal candidates. We call the anchor locations "anchor maps" as in Fig. 1. (The highlighted points actually locate at multiple channels of score maps and we only show one channel for visual simplicity.) Then proposals with high scores will be sent to R-CNN for further classification to filter out false positives and keep confident boxes as final detection results. During the procedure we record the source anchor location of each detected box and keep the specific location at anchor maps as peaks. Later the PRM phase will start from the peaks at score maps, which is interpreted as a random walker procedure from the peaks to the bottom layer in [14]. Assuming that $U$ and $V$ are the input and output feature map of a convolution layer in the forward process, whose filter size is $s \times h \times w$. The visiting probability of the random walker or the correlation between spatial locations $U_{ijk}$ and $V_{pqt}$ during PRM can be formulated by

$$P\left(U_{ijk}\right) = \sum\nolimits_{p=i-\frac{s}{2}}^{i+\frac{s}{2}} \sum\nolimits_{q=j-\frac{h}{2}}^{j+\frac{h}{2}} \sum\nolimits_{t=k-\frac{w}{2}}^{k+\frac{w}{2}} P\left(U_{ijk}|V_{pqt}\right) \times P\left(V_{pqt}\right) \tag{1}$$

where the conditional probability is

$$P\left(U_{ijk}|V_{pqt}\right) = Z_{pqt} \times \widehat{U}_{ijk} W^{+}_{(i-p)(j-q)(k-t)} \tag{2}$$

$\widehat{U}_{ijk}$ denotes the activation value at location $(i, j, k)$ of $U$ during the forward process. $W^+$ means that we only reserve the positive weights of filter and $Z_{pqt}$ is normalization factor to ensure $\sum_{i,j,k} P\left(U_{ijk}|V_{pqt}\right) = 1$. Note that the PRM can be realized using normal gradient back-propagation during inference and does not require any extra conditions or constraints for network training.

***An Advanced 2D Otsu for Instance Segmentation.*** Figure 2 shows several examples of visual cues produced by PRM, from which we can tell the contour or boundaries of soma instance. But the PRM is not perfect enough as instance mask for two defects: (1) the regions highlighted as the most discriminative parts are usually not complete and may be broken; (2) other instances may also appear around targets in PRM, so only utilizing PRM can not remove such false regions. Hence, we propose a thresholding method for segmentation that utilizes the PRM results but in addition utilizes the grayscale information. Our method is an advanced 2D Otsu algorithm, but different from the traditional 2D Otsu where the second-dimension is using manually crafted features [12], we use the visual cues extracted from the deep learning-based detection network as the second dimension, which provides complementary information to the local intensity. In order to balance the weights between PRM and intensity, we rescale both into the same dynamic range, and we design a 2D oblique segmentation on the 2D histogram to leverage the complementary information. Specifically, assuming $G$ and $P$ are the intensity and PRM values of voxels inside one detected box, which constitute the two axes of the 2D histogram. Then the thresholding acts with an oblique decision boundary: $\text{sign}(G + k \times P - b)$ where $k$ and $b$ are the slope and bias. And $k$ is set as 1 for the same sale of $G$ and $P$, and $b$ is searched in a recursion way aiming to maximize the between-class variance. Once the decision boundary is fixed, the segmentation can be accomplished by the thresholding.

## 3 Experiments and Results

We conduct experiments on two volumetric biomedical datasets, both are optical microscopy images: mouse brain soma dataset collected by ourselves and nuclei of HL60 cells [6,10]. Our method uses merely bounding box labels for training and evaluates instance segmentation performance on voxel-wise labeled test set. Considering that we aim to tackle both volumetric data and learning without voxel-wise annotation problems for instance segmentation task, so we select several competitive methods satisfying both conditions as comparison.

***Mouse Brain Soma Data.*** Our mouse brain soma data is acquired by fluorescence staining and confocal microscopy imaging techniques, whose resolution is high enough to distinguish each neuron. For the training set, we label soma with inscribed sphere and get bounding box labels by extending the globules.

**Table 1.** Results of average precision with three volumetric mask IoU thresholds on our mouse brain soma dataset. All of these methods do not require voxel-wise labeling.

| Method | Instance segmentation AP | | |
|---|---|---|---|
| | IoU 0.3 | IoU 0.4 | IoU 0.5 |
| NeuroGPS [7] | 0.4965 | 0.3960 | 0.2705 |
| DSN [2] | 0.5459 | 0.4236 | 0.2512 |
| Detection+1D Otsu | 0.6563 | 0.5077 | 0.3904 |
| Detection+2D Otsu (w/o PRM) | 0.6741 | 0.5333 | 0.3992 |
| Detection+2D Otsu (w/ PRM) | **0.7024** | **0.5864** | **0.4253** |

**Table 2.** Results on the HL60 cells dataset in terms of F1 score (the box IoU threshold is 0.4). Det denotes detection.

| Method | Voxel-wise labeling | Detection F1 | | | Segmentation F1 | | |
|---|---|---|---|---|---|---|---|
| | | Track1 | Track2 | Mean | Track1 | Track2 | Mean |
| Mask R-CNN [13] | 20% | 0.9967 | **0.9599** | **0.9783** | 0.9416 | 0.8437 | 0.8927 |
| VoxResNet [13] | 4/13 | 0.9965 | 0.9221 | 0.9593 | 0.9610 | 0.7873 | 0.8742 |
| Det+1D Otsu | 0 | **0.9970** | 0.9545 | 0.9758 | 0.7708 | 0.5668 | 0.6688 |
| Det+2D Otsu (w/o PRM) | 0 | | | | 0.8792 | 0.6230 | 0.7511 |
| Det+2D Otsu (w/ PRM) | 0 | | | | **0.8902** | **0.7399** | **0.8151** |

All images are processed into the same size of $128 \times 256 \times 256$ and saved as 16bit images whose physical resolution is $1\mu\text{m}^3$/voxel. Figure 2 shows several image examples. We have 3000 and 228 images for training and testing respectively. Considering the small soma targets and the memory limitation for CNN to handle volumetric data, we set feature stride as 4, input size as $64\times256\times256$ and batch size as 4 on two GeForce GTX 1080Ti's. Note that we can use such big input size thanks to the compact network structure. During inference, the detector firstly outputs boxes and corresponding scores. Afterwards the visual cues are produced by back-propagation for every box. Then the thresholding is completed off-line.

For there is no existing report on our soma data, we compare our method with two advanced methods including an optimization-based method NeuroGPS [7] and a learning-based semantic segmentation method DSN using course mask label [2], both of which do not need voxel-wise label. NeuroGPS is designed for neuron images like our soma data, which aims to find the most appropriate center coordinates and its radius for soma, so we regard the detected solid globules as instance masks. For DSN we find all the connected components as instance segmentation results. And for both baselines those too small masks are excluded to balance precision and recall. In addition, to verify the advantage of our 2D Otsu algorithm, we also check the results of simple but still powerful thresholding methods including 1D Otsu with grayscale only and 2D Otsu whose second dimension is Gaussian filtered grayscale. We evaluate the performance using

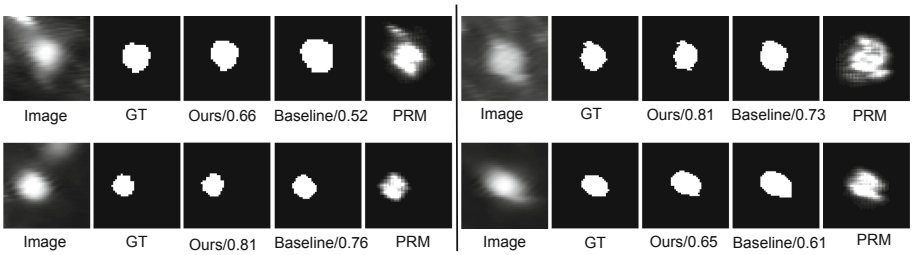


**Fig. 3.** Our proposed 2D Otsu with PRM produces segmentation results with more fine details, where IoU is calculated between segmentation result and ground-truth (GT).

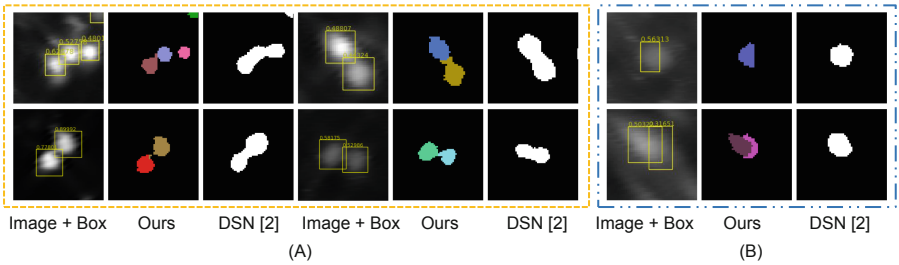


**Fig. 4.** (A) Detection helps identify instance individual in crowded-instance cases. (B) Failure cases. Best viewed in color.

Average Precision (AP) with three different IoU thresholds for volumetric masks, and results are shown in Table 1. Our approach achieves the best performance for all the metrics. And the ablation study shown in latter three rows indicates that with PRM as visual cues the segmentation AP can gain up to 5.3%.

***Nuclei of HL60 Cells.*** We also apply our method to nuclei of HL60 cells [6,10], a synthetic dataset that contains two tracks and has full voxel-wise annotation for all instances. Note that this dataset is synthetic and the existing reported methods still more or less require voxel-wise labeling, thus the comparison is not fair. However, this is the only public dataset for our task and we conduct experiment on this dataset to verify the generality of our method. We follow the training/test split in [13]. Because the cells in this dataset are larger than those in our soma data, we use feature stride as 8. Also considering that this dataset has voxel-wise annotation and the related work from [13] is more competitive, we compare both detection and instance segmentation performance with the results in [13], using the same evaluation metric: F1 score with IoU threshold of 0.4. We also perform ablation study to verify our PRM-based 2D Otsu algorithm. Table 2 presents F1 scores of all methods. Our approach can achieve comparable detection performance with [13]. For instance segmentation, our method performs not as well as [13] because we use no voxel-wise label in

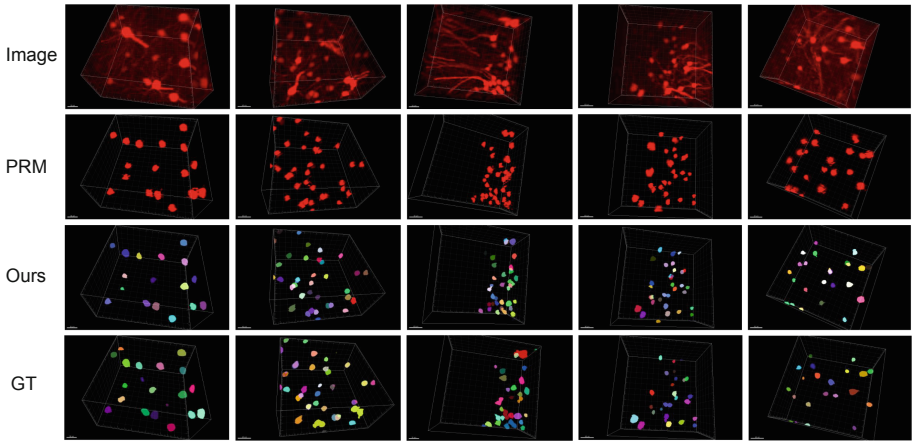


**Fig. 5.** Visualized PRM and instance segmentation results, where different instances are marked with random colors. Best viewed in color.

training, and for fairness the F1 scores in bold shows the best ones without any voxel-wise label. However, as this dataset is synthetic, it may not reveal the real-world cases.

***Visualization Results.*** To better understand the effect of PRM and detection and visually evaluate our method, here we illustrate three groups of soma images in both 2D and 3D views. Figure 3 shows how PRM benefits thresholding. These results suggest that with PRM as a guidance the segmented mask has a closer appearance to the ground truth, especially in fine details. Although the visual cues from PRM might be not complete, the discriminative regions serve as complementary information to intensity and improve the mask contour. Examples in Fig. 4 reflect the influence of detection on instance segmentation. Group A shows that detected boxes help identify soma instances in dense-soma case even when they touch each other. While in some low-contrast or cropped regions, the segmentation may be not good enough for the boxes not precisely detected. Figure 5 illustrates some results in 3D view. We can see that our method can detect and segment soma in diverse intensity, density, shapes as well as from complex background, and the appearance is quite close to the ground truth.

## 4 Conclusion

In this paper, we propose a weakly supervised instance segmentation method for volumetric biomedical images not requiring any voxel-wise label. The network can be trained as a simple detector with bounding boxes only. And instance segmentation can be accomplished by PRM combined with an advanced thresholding algorithm. We design experiments on two datasets and results demonstrate

the efficiency of the proposed method. Our approach can save considerable labeling efforts and has potential to be applied to other related segmentation tasks.

**Acknowledgements.** This work was supported by the Natural Science Foundation of China under Grant 91732304, and by the Fundamental Research Funds for the Central Universities under Grant WK2380000002.

## References

1. Çiçek, Ö., et al.: 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In: MICCAI. pp. 424–432 (2016)
2. Dong, M., et al.: 3D CNN-based soma segmentation from brain images at single-neuron resolution. In: ICIP. pp. 126–130 (2018)
3. Dou, Q., et al.: 3D deeply supervised network for automatic liver segmentation from CT volumes. In: MICCAI. pp. 149–157 (2016)
4. He, K., et al.: Mask R-CNN. In: ICCV. pp. 2961–2969 (2017)
5. Khoreva, A., et al.: Simple does it: Weakly supervised instance and semantic segmentation. In: CVPR. pp. 876–885 (2017)
6. Maška, M., et al.: A benchmark for comparison of cell tracking algorithms. Bioinformatics **30**(11), 1609–1617 (2014)
7. Quan, T., et al.: NeuroGPS: Automated localization of neurons for brain circuits using L1 minimization model. Scientific Reports 3, Article No. 1414 (2013)
8. Ren, S., et al.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: NIPS. pp. 91–99 (2015)
9. Simonyan, K., et al.: Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)
10. Ulman, V., et al.: An objective comparison of cell-tracking algorithms. Nature Methods 14(12), Article No. 1141 (2017)
11. Yang, L., et al.: Suggestive annotation: A deep active learning framework for biomedical image segmentation. In: MICCAI. pp. 399–407 (2017)
12. Zhang, J., et al.: Image segmentation based on 2D Otsu method with histogram analysis. CASCON. **6**, 105–108 (2008)
13. Zhao, Z., et al.: Deep learning based instance segmentation in 3D biomedical images using weak annotation. In: MICCAI. pp. 352–360 (2018)
14. Zhou, Y., et al.: Weakly supervised instance segmentation using class peak response. In: CVPR. pp. 3791–3800 (2018)