doi:10.1017/S143192762002053X
Microsc. Microanal. 26 (Suppl 2), 2020

Enforcing Prediction Consistency Across Orthogonal Planes Significantly Improves Segmentation of FIB-SEM Image Volumes by 2D Neural Networks.

Ryan Conrad, Hanbin Lee and Kedar Narayan

National Cancer Institute, NIH & Frederick National Laboratory for Cancer Research, Frederick, Maryland, United States

Volume electron microscopy (vEM) is transforming cell biology by generating high-resolution 3D reconstructions of large biological samples. However, segmentation of specific features such as mitochondria from highly heterogenous image volumes remains a bottleneck – even powerful deep learning (DL) approaches reveal various limitations and artifacts [1]. Within vEM technologies, focused ion beam scanning electron microscopy (FIB-SEM) can yield isotropic-voxel data where information in orthogonal planes (*xy*, *xz*, *yz*) is essentially interchangeable; here, we exploit this to develop a two-step DL algorithm for segmentation. First, we train a DL model to segment specific features in 2D image slices. Crucially, to incorporate 3D context into the predicted segmentation of a target volume, we run inference over *xy*, *xz* and *yz* planes, and average the results at each voxel, a procedure we call "ortho-plane inference". In the second step, we use the target volume and predicted segmentation to train a new 2D model in a weakly supervised setting with "bootstrapping". Bootstrapping enforces prediction consistency between adjacent voxels of the same object regardless of viewing orientation. This two-step algorithm results in a 23% increase in Intersection-over-Union (IoU) over the best case scenarios for ortho-plane inference without bootstrapping and a 35% IoU increase over "2D stack" inference (Fig 2). Sampling 3D volumes while staying in a 2D regime makes this approach nimble and thus well suited to vEM researchers with limited image and compute resources.

In DL, 2D models are the most memory and data efficient option for 3D image segmentation. A volume containing 100 cubic voxels that would be a single example for a 3D model generates 300 examples for a 2D model after slicing along the principal axes. This 2D model would also have roughly 3x fewer parameters and could readily be initialized with weights pretrained on ImageNet. To address the chief disadvantage of working in 2D, i.e. the loss of valuable 3D context, we incorporate 3D information through ortho-plane inference. This inference strategy results in improved performance but is diminished by two key weaknesses in the model: a susceptibility to small changes in object appearance between adjacent image slices and to larger changes between orthogonal slices. Common examples of these errors are the "stacked pancake" artifact, familiar to researchers in the vEM field, and "cross-hatching" patterns, shown in Figure 2. In this advance, we train a second dataset-specific neural network to learn the noise patterns associated with these errors and eliminate them.

During training of this second neural network we modify our target labels to be a combination of the model's hardened predictions (0s or 1s) and the noisy labelmap created by ortho-plane inference. This application of bootstrapping enforces prediction consistency, which, in turn, amplifies the signal present in noisy labels [2]. The value of generated labels is given by:

$$q_i = \beta y_i + (1-\beta)\mathbb{1}_{p_i > 0.5}$$

Where y_i and p_i are the noisy label and model prediction at pixel i, respectively, and β is a hyperparameter with value between 0 and 1, following [2] we set β=0.8. Our training criterion is then the dice loss between the generated labels and the model's soft prediction confidence [3]. This particular learning setting is considered weakly supervised.

We evaluate our method by applying a supervised model, previously trained on segmented mitochondria from a small labeled FIB-SEM sub-volume, to a large target volume from a separate experiment and a substantially different cell sample. The supervised model was trained on a manually annotated 224 voxel cube (672 slices along orthogonal axes). We trained a DeepLabV3 model with ResNet34 backbone, pretrained on ImageNet, for 10000 iterations using dice loss and the OneCycle learning rate policy with Adam optimizer, learning rate 0.001, batch size of 64, and dropout of 0.5 after the atrous spatial pyramid pooling module [4][5]. To alleviate overfitting, we froze the weights in the model backbone below the fourth ResNet block. Data augmentations included random resized crops, horizontal and vertical flips, brightness and contrast adjustment, and Gaussian noise. To create the noisy labelmap for the bootstrapping step, we ran ortho-plane inference on the target volume and set the confidence threshold at 0.1. For the weakly supervised step, we used the same architecture, hyperparameters, and inference strategy but with the confidence threshold set at 0.5. The workflow is shown in Figure 1.

The noisy and final segmentation results are shown in Figure 2. We compute each prediction's IoU with a manually labeled ground truth. Running inference strictly on the imaging plane stack, which is a common approach, reaches a best case IoU of 0.48. By accessing orthogonal views with our ortho-plane inference, we achieve a best case IoU of 0.53 (a 9% increase). Critically, by further incorporating bootstrapping on top of ortho-plane inference, we boost the IoU by an additional 23%, achieving scores of up to 0.65. We also observe more consistent performance over a range of confidence thresholds, which indicates better agreement between predictions made on orthogonal planes. Although the results are promising, it should be noted that this approach cannot fix systematic errors; vesicles and lipid bodies labeled incorrectly, but consistently, in the original prediction will be retained. Additional postprocessing before bootstrapping could help to further boost performance. Application of this method to other vEM datasets that do not yield isotropic voxels is possible but may require more training data as learned features are unlikely to transfer as well across different image planes. Overall, the simplicity, effectiveness and low cost of our method makes it a useful tool for use as it is and a template for further improvements in addressing the vEM segmentation bottleneck.

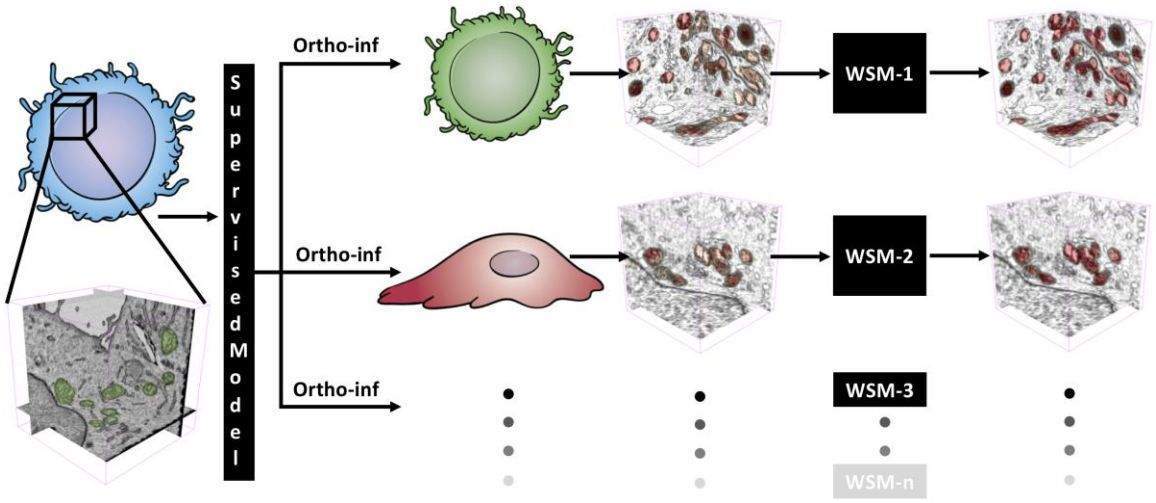

Figure 1. Schematic of our workflow enforcing prediction consistency across orthogonal planes. We train a supervised model on a small labeled 3D ROI from a FIB-SEM reconstruction and run ortho-plane inferences (ortho-inf) on larger, unrelated target volumes. The initial noisy outputs of these inferences are then used to train weakly supervised models (WSM) for each target volume with a process called bootstrapping. The resulting dedicated WSMs show significantly improved performances and can be run on any number of FIB-SEM datasets.

https://doi.org/10.1017/S143192762002053X Published online by Cambridge University Press

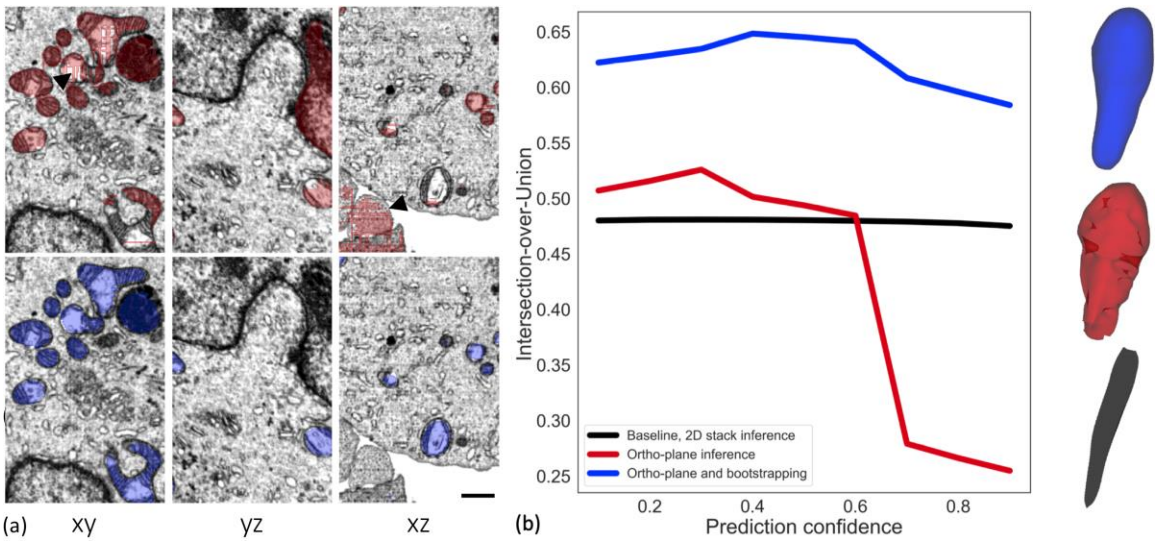

Figure 2. (a) Results from mitochondrial segmentation from running ortho-plane inference without (top) and with (bottom) bootstrapping. Arbitrary slices along three principal axes from a FIB-SEM cellular volume reconstruction are shown, with examples of the "cross-hatching" error indicated (arrowheads). Scale bar, 1 μm (b) Quantitative evaluation of IoU metric from this dataset and volume rendering of a representative mitochondrion after our advance (ortho-plane inference and bootstrapping, blue) as compared to no bootstrapping (red) and 2D stack inference (black).

References

[1] Xiao, C., Chen, X., Li, W., Li, L., Wang, L., Xie, Q., & Han, H. (2018). Automatic Mitochondria Segmentation for EM Data Using a 3D Supervised Convolutional Network. *Frontiers in Neuroanatomy* ***12***, 92.

[2] Reed, S. E., Lee, H., Anguelov, D., Szegedy, C., Erhan, D., & Rabinovich, A. (2014). Training Deep Learning Neural Networks on Noisy Labels with Bootstrapping. https://arxiv.org/abs/1412.6596

[3] Milletari, F., Navab, N., & Ahmadi, S.-A. (2016). V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. *Proceedings - 2016 4th International Conference on 3D Vision,* ***3DV 2016***, 565–571. http://arxiv.org/abs/1606.04797

[4] Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. *Lecture Notes in Computer Science*, ***11211 LNCS***, 833–851.

[5] Smith, L. N. (2018). A disciplined approach to neural network hyper-parameters: Part 1 -- learning rate, batch size, momentum, and weight decay. http://arxiv.org/abs/1803.09820

We thank Adam Harned for acquiring the FIB-SEM datasets used throughout this work and Dr. Stanley Lipkowitz for providing the cell samples.This project has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, under Contract No. 75N91019D00024. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.