frontiers | Frontiers in Bioinformatics

TYPE Original Research
PUBLISHED 15 December 2023
DOI 10.3389/fbinf.2023.1308707

OPEN ACCESS

EDITED BY
Jia-Ren Lin,
Harvard Medical School, United States

REVIEWED BY
Xun Wang,
China University of Petroleum (East China), China
Beth A. Cimini,
Broad Institute, United States

*CORRESPONDENCE
Lucas Pagano,
paganol@ohsu.edu

RECEIVED 06 October 2023
ACCEPTED 29 November 2023
PUBLISHED 15 December 2023

CITATION
Pagano L, Thibault G, Bousselham W, Riesterer JL, Song X and Gray JW (2023), Efficient semi-supervised semantic segmentation of electron microscopy cancer images with sparse annotations. *Front. Bioinform.* 3:1308707. doi: 10.3389/fbinf.2023.1308707

Efficient semi-supervised semantic segmentation of electron microscopy cancer images with sparse annotations

Lucas Pagano[1,2]*, Guillaume Thibault[1], Walid Bousselham[1], Jessica L. Riesterer[1,2], Xubo Song[2,3] and Joe W. Gray[1]

[1]Department of Biomedical Engineering, Oregon Health and Science University, Portland, OR, United States, [2]Knight Cancer Institute, Oregon Health and Science University, Portland, OR, United States, [3]Department of Medical Informatics and Clinical Epidemiology at Oregon Health and Science University, Portland, OR, United States

Electron microscopy (EM) enables imaging at a resolution of nanometers and can shed light on how cancer evolves to develop resistance to therapy. Acquiring these images has become a routine task.However, analyzing them is now a bottleneck, as manual structure identification is very time-consuming and can take up to several months for a single sample. Deep learning approaches offer a suitable solution to speed up the analysis. In this work, we present a study of several state-of-the-art deep learning models for the task of segmenting nuclei and nucleoli in volumes from tumor biopsies. We compared previous results obtained with the ResUNet architecture to the more recent UNet++, FracTALResNet, SenFormer, and CEECNet models. In addition, we explored the utilization of unlabeled images through semi-supervised learning with Cross Pseudo Supervision. We have trained and evaluated all of the models on sparse manual labels from three fully annotated in-house datasets that we have made available on demand, demonstrating improvements in terms of 3D Dice score. From the analysis of these results, we drew conclusions on the relative gains of using more complex models, and semi-supervised learning as well as the next steps for the mitigation of the manual segmentation bottleneck.

KEYWORDS

deep learning, semi-supervised, semantic segmentation, electron microscopy, vEM, sparse labels

1 Introduction

Recent advances in cancer nanomedicine have made cancer treatment safer and more effective (Shi et al., 2017). Nanotechnology has elucidated interactions between tumor cells and their microenvironment showing key factors in cancer behavior and responses to treatment (Hirata and Sahai, 2017; Tanaka and Kano, 2018). Gaining a deeper understanding of the underlying mechanisms taking place during such interactions will help us understand how cancer grows and develops drug resistance, and ultimately help us find new, efficient, and safe therapeutic strategies aimed at disrupting cancer development (Baghban et al., 2020).

To do this, high-resolution information collected from the cellular components at a nanometer scale using focused ion beam-scanning electron microscopy (FIB-SEM) is especially useful as it provides volumes of serially-collected 2D SEM images, creating

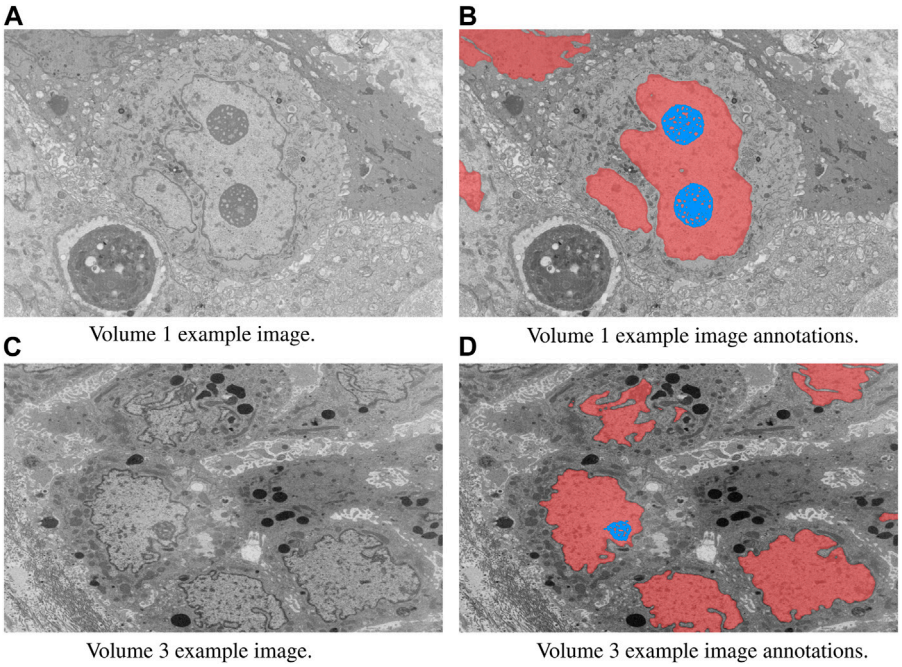

FIGURE 1
Example image slices and corresponding ground truth annotations from Volumes 1 and 3. The horizontal image width is equal to 25 μm. Nuclei are in red and nucleoli in blue. **(A)** Volume 1 example image. **(B)** Volume 1 example image annotations. **(C)** Volume 3 example image. **(D)** Volume 3 example image annotations.

volume EM (vEM) image stacks, and allowing access to 3D information from tissues (Giannuzzi and Stevie, 2005). This fully automated protocol avoids artifacts associated with serial microtomies and enables voxels to be isotropic, thus yielding a similar image quality in all dimensions, beneficial for feature recognition and context within the volume (Bushby et al., 2011).

These advantageous features have made SEM desirable for use in clinical programs. However, the analysis-limiting step is the extraction of meaningful features, starting with the segmentation of cellular components present in these images. This is currently done by human experts through hand annotation. It is a tedious and time-consuming task, making it unsuitable for medical applications and decisions where time is a critical factor. To overcome this limitation and fully make use of FIB-SEM in a clinical setting, the development of automated and robust models is critical to speeding up this task (Perez et al., 2014).

Segmenting images acquired via FIB-SEM is a difficult problem. Indeed, these images differ considerably from natural ones (images representing what human beings would observe in the real world), and even from other microscopy techniques such as fluorescence microscopy, due to increased noise, different collection resolution, and the reduced number of image channels. EM images are single-channel (grayscale) and tend to have limited contrast between objects of interest and background (Karabğ et al., 2020). Furthermore, the ultrastructure of tumor cells and their microenvironment vary from those of normal cells (Nunney et al., 2015), and EM analysis methods can be tissue type dependent; most current methods have been developed for neural images (Taha and Hanbury, 2015; Zhang et al., 2021). Therefore, segmentation methods designed to assist other microscopy modalities or other tissue types cannot be applied to ultrastructure segmentation of cancer cells imaged by EM.

We expanded on previous work from Machireddy et al. (2021), where authors showed that a sparsely manually annotated dataset, typically around 1% of the image stack, was sufficient to train models to segment the whole volume. While state-of-the-art in semantic segmentation has been dominated by attention-based models for natural images (Jain et al., 2021), convolutional architectures remain mainstream with EM data, and were used in Machireddy et al. (2021), and the companion paper within this journal volume. In this paper, we compared architectures as well as training frameworks to find the most suitable one for the task of semantic segmentation in the aforementioned specific context of FIB-SEM images. By optimizing the learning process, we expected to improve overall segmentation results and minimize the manual annotation bottleneck by reducing the number of manually labeled images needed for training.

In this study, we focused on the segmentation of nuclei and nucleoli in vEM image stacks acquired from human tumor samples, as both are commonly used as cancer cell identifiers (Zink et al., 2004) and have emerged as promising therapeutic targets for cancer treatment (Lindström et al., 2018). Segmenting both structures accurately has thus proven essential. We evaluated the selected models on FIB-SEM images of three longitudinal tissue biopsy datasets that are available on demand as part of the Human Tumor Atlas Network (HTAN). A quick visualization of the data and end results can be found in Figures 1, 2.

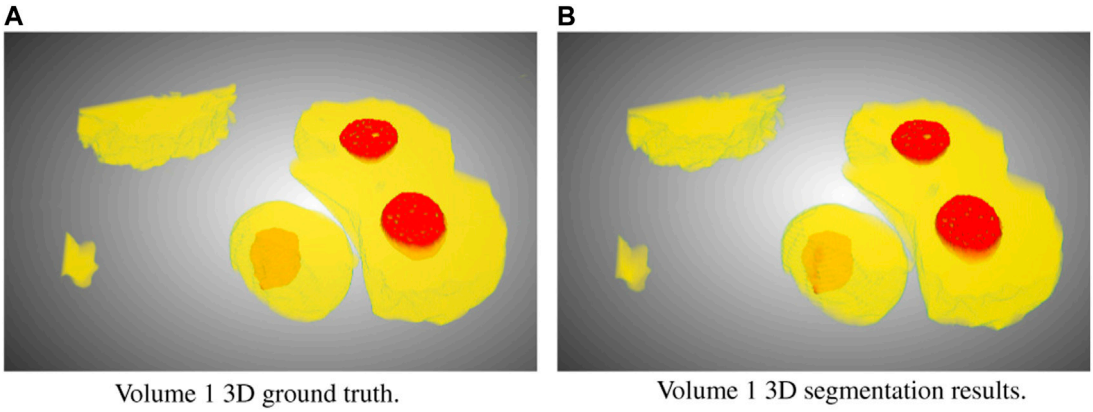

FIGURE 2
Ground truth **(A)** and segmentation by SSL-UNet++-CutMix **(B)** 3D visualizations for Volume 1. Nuclei are in yellow and nucleoli in red.

2 Materials and methods

2.1 Image collection and preprocessing

2.1.1 Dataset collection

In a sanctioned observational study approved by the institutional review board, we obtained volumes 2 and 3 at two distinct time points during cancer treatment from a patient diagnosed with metastatic ER + breast ductal carcinoma. Additionally, volume 1 was procured from a patient with pancreatic ductal adenocarcinoma. All biopsies underwent acquisition and analysis in accordance with the OHSU IRB-approved Molecular Mechanisms of Tumor Evolution and Resistance to Therapy protocol (IRB#16113). The HTAN sample IDs for volumes 2 and 3 are HTA9-1 bx1 and HTA9-1 bx2, respectively. Informed written consent was secured from all subjects.

The samples were preserved using Karnovsky's fixative [2.0% PFA, 2.5% Glutaraldehyde (Karnovsky, 1964)], post-fixed utilizing an OsO4-TCH-OsO4 staining protocol, and embedded in Epon resin (Riesterer et al., 2019). Post-fixation staining involved binding heavy metals to lipid-rich membranes to enhance contrast in electron microscopy imaging. To achieve high-resolution, charge-free, high-contrast, and low-noise images, a conductive coating with an 8 nm-thick layer of carbon was applied. The imaging process was carried out using a FEI Helios NanoLab 660 DualBeam™ FIB-SEM to collect high-resolution 3D volumes of the resin-embedded blocks.

Targeted volumes were obtained through the use of a Ga + FIB source, sequentially slicing a few nanometers from the sample to expose a fresh surface for subsequent imaging. The slicing/imaging cycle was automated through the FEI AutoSlice and View™ software extended package, with image collection during 3D data acquisition utilizing the in-column detector (ICD). The isotropic resolution for imaging metastatic breast cancer and primary pancreatic tissues was 4 and 6 nm, respectively.

2.1.2 Image preprocessing

Following data acquisition, images within the stack underwent translational alignment in the XY-plane using an internally developed stochastic version of TurboReg affine transformation (Thevenaz et al., 1998). During the alignment step, zero-padding was applied to the images to maintain a consistent size, after which they were cropped to generate the final 3D image volumes. The registration and edge cropping procedures resulted in a final resolution of 6065 × 3976 × 792 for volume 1, 5634 × 1912 × 757 for volume 2, and 5728 × 3511 × 2526 for volume 3.

At times, the brightness of certain images in the stack exhibited variability, introducing complexity and rendering segmentation more challenging. To ensure uniformity across the stack and mitigate image complexity, histogram equalization was implemented.

2.2 Training and evaluation

Previously, we trained a model for each dataset using a subset of manually labeled images spaced evenly along the volumes and evaluated on remaining unlabeled images. We reported results on using 7, 10, 15, and 25 training images on all datasets, which represents between 0.3% and 3.3% of all image slices depending on the dataset. Volume 1, 2, and 3 have a total of 792, 757, and 2526 slices respectively; all slides except for the training ones were used for producing evaluation Tables 3–6. As these EM images had large dimensions (typically around 6000 × 4000 pixels), they were cropped to 512 × 512 tiles. We followed the same procedure as (Machireddy et al., 2021) of extracting tiles of size 2048 × 2048 and downsampling them to 512 × 512 as a way to capture larger spatial context. We applied standard random flip and rotation data augmentations. When training with nucleoli, as they account for a small area in the total image, taking random crops effectively resulted in most crops being empty, and models collapsing to the prediction of background. To address this issue, we ensured that more than 99% of crops in a batch contain nucleoli.

Moreover, we selected the Dice score as an evaluation metric because it can be seen as a harmonic mean of precision and recall, and is fitted when dealing with imbalanced class settings, which is our case. However, we also report the 3D Dice score rather than the average of individual Dice scores across all slices as reported in Machireddy et al. (2021). Indeed, we found the latter to be biased

TABLE 1 Number of parameters and training times for one volume. SSL-trained models require double the number of parameters because two models are trained at the same time.

	UNet++	FracTALResNet	CEECNet	SenFormer	SSL-ResUNet	SSL-UNet++
# of parameters	26,072,337	18,199,919	58,964,079	163,100,906	4,283,201 * 2	21,954,705 * 2
Average training time	~48 h	~70 h	~90 h	~144 h	~60 h	~64 h

towards giving more importance to slices with fewer foreground pixels, while the former effectively reflects the captured percentage of the target structure. For the sake of comparison, we reported the averaged version in our results section in addition to the 3D Dice scores. We recommend however to use the latter. The 3D Dice and average dice are more precisely defined as follows:

$$3D\ Dice = Dice(Predicted\ Volume, Ground\ Truth\ Volume)$$

$$Average\ Dice = \frac{Sum(Dice(Predicted\ Slice, Ground\ Truth\ Slice))}{Number\ of\ slices\ in\ volume}$$

All models were trained and evaluated on one NVIDIA V100 GPU, as we strongly believe we should keep our clinical end goal in mind and aim to reflect image analysis capabilities available to teams with reasonable computational power. To this end, we also report training times and the number of parameters in Table 1.

2.3 Fully-supervised framework

2.3.1 ResUNet

We used previous work from Machireddy et al. (2021) as the baseline for nuclei and nucleoli segmentation. The model used was a Residual U-Net (ResUNet) (He et al., 2015), a simple yet robust fully convolutional encoder-decoder network. U-net and its variants are the most prominent architectures for image segmentation, as the residual connections solve the gradient vanishing problem faced when working with very deep models (Glorot and Bengio, 2010), while the different levels allow feature refinement at different scales. These features have made U-nets widely used in many computer vision problems, including analysis of medical data (Siddique et al., 2020; Su et al., 2021).

2.3.2 UNet++

UNet++ (Zhou et al., 2018) was the first model we decided to compare to the baseline. Our motivation to use this model came from the fact that it was heavily inspired by ResUNet, and was especially designed for medical-like image segmentation. The major differences from ResUNet are the presence of dense convolution blocks on the skip connections and deep supervision losses. Dense convolution blocks aim at reducing the semantic gap between the encoder and decoder, while deep supervision loss enables the model to be accurate (by averaging outputs from all segmentation branches) and fast (by selecting one of the segmentation maps as output). In our work, as we were primarily focused on accuracy, we used the average of all branches. We used the implementation available in the *Segmentation Models* Python library[1] with ResNet34 as the encoder backbone, and the soft Dice loss (DL) function which is commonly used in semantic segmentation for images when background and foreground classes are imbalanced and is defined as follows for a ground truth y and prediction $\hat{p}$:

$$DL(y, \hat{p}) = 1 - \frac{2y\hat{p}}{y + \hat{p}} \quad (1)$$

2.3.3 FracTALResNet

FracTALResNet (Diakogiannis et al., 2021) was also used for comparison. While the original model presented is designed for the task of semantic change detection, it can be adapted for semantic segmentation, and such architecture is in fact available in the authors' official implementation[2]. It was heavily inspired by ResUNet as well, but makes use of a multi-head attention layer (FracTAL block). It also makes use of boundaries and distance maps calculated from the segmentation masks in order to improve performances, but at the cost of both memory and computational time during training. It is trained using the Fractal Tanimoto similarity measure.

2.3.4 CEECNet

CEECNet was also introduced in Diakogiannis et al. (2021) and for the same purpose as FracTALResNet but managed to achieve state-of-the-art performances by focusing on context. Indeed, the CEECNet block stands for Compress-Expand Expand-Compress and is comprised of two branches. The first branch (CE block) processes a view of the input in lower resolution, while the second branch (EC block) treats a view in higher spatial resolution. The motivation behind using this model came from the fact that, as described in Section 2.2, feeding more context by down-sampling to a lower resolution is beneficial to segmentation accuracy. Since the core block of CEECNet is based on the compress and expand operations, we believed this network would be able to leverage contextual information in order to achieve better segmentation performances. Similar to FracTALResNet, it was trained with the Tanimoto similarity measure and needs computed boundaries and distance maps.

2.3.5 SenFormer

SenFormer (Bousselham et al., 2021) (Efficient Self-Ensemble Framework for Semantic Segmentation) was the last fully-supervised method tested. It is a newly developed ensemble approach for the task of semantic segmentation that makes use of transformers in the decoders and the Feature Pyramid Network (FPN) backbone. Our motivation behind using this model came from the fact that it is almost purely attention-based, which by definition adds spatial context to the segmentation.

1 https://github.com/qubvel/segmentation_models.pytorch

2 https://github.com/feevos/ceecnet

TABLE 2 Average, standard deviation, and average rank for dice score over all volumes.

	ResUNet	UNet++	FracTALResNet	CEECNet	Senformer
Average	0.9200	**0.9270**	0.9250	0.9036	0.9146
Standard deviation	0.0107	0.0083	**0.0072**	0.137	0.0099
Average rank	2.625	**2.6042**	3.1875	3.54	3.0417

Bold values represent best results.

TABLE 3 Nuclei segmentation dice scores. Columns labeled 7, 10, 15, and 25 in the second line represent the number of training images for each volume.

	Volume 1				Volume 2				Volume 3			
	7	10	15	25	7	10	15	25	7	10	15	25
ResUNet	0.9805	0.9846	0.9846	0.9847	0.9845	0.9875	0.9878	0.9933	0.9597	0.9737	0.9738	0.9745
UNet++	0.9746	0.9791	0.9801	0.9844	0.988	0.9908	0.9922	0.993	0.9606	0.9625	0.9705	0.985
FracTALResNet	0.9724	0.9796	0.9817	0.9885	0.9756	0.9836	0.9871	0.9887	0.9655	0.9698	0.976	0.9825
CEECNet	0.9702	0.9688	0.9825	0.9742	0.9847	0.9894	0.9928	0.9948	0.9457	0.9499	0.9509	0.98
Senformer	0.9821	0.9851	0.9877	0.9897	0.9835	0.987	0.9898	0.9927	0.971	0.9722	0.979	0.9858
SSL-ResUNet	0.988	0.9889	0.9882	0.9902	0.9938	0.9931	0.9941	0.9958	0.9703	0.9702	0.977	0.9822
SSL-ResUNet-CutMix	0.9892	0.9898	0.9903	0.9912	**0.9951**	**0.9952**	**0.9954**	0.9957	0.9726	0.9747	0.9799	0.9822
SSL-UNet++-CutMix	**0.9903**	**0.9910**	**0.9912**	**0.9923**	**0.9951**	**0.9952**	**0.9954**	**0.9959**	**0.9804**	**0.9839**	**0.9859**	**0.9872**

Bold values represent best results.

TABLE 4 Nucleoli segmentation dice scores. Columns labeled 7, 10, 15, and 25 in the second line represent the number of training images for each volume.

	Volume 1				Volume 2				Volume 3			
	7	10	15	25	7	10	15	25	7	10	15	25
ResUNet	0.9686	0.9733	0.9664	0.9712	0.8811	0.9019	0.9108	0.9182	0.7054	0.7014	0.6906	0.7166
UNet++	0.9576	0.9624	0.9652	0.9671	0.8957	0.9168	0.9139	0.9216	0.638	0.7321	0.7893	0.8282
FracTALResNet	0.9662	0.9613	0.9464	0.9671	0.8809	0.8778	0.8775	0.9237	0.6892	0.7547	0.7677	0.8307
CEECNet	0.9779	0.9775	0.9772	**0.9797**	0.7615	0.8608	0.8339	0.8565	0.586	0.6567	0.7321	0.8036
Senformer	0.9353	0.9384	0.9413	0.9442	0.8385	0.8594	0.8818	0.907	0.6695	0.6678	0.7538	0.8084
SSL-ResUNet	0.9775	**0.9783**	0.9767	0.9782	0.9007	0.9196	0.9344	0.9401	0.6714	0.7447	0.8041	0.8176
SSL-ResUNet-CutMix	**0.9781**	0.9782	**0.9787**	0.9791	0.9193	0.9218	0.9339	0.9440	0.7763	0.8013	0.8159	0.8266
SSL-UNet++-CutMix	0.9777	**0.9783**	0.9781	0.9793	**0.9292**	**0.9256**	**0.9349**	**0.9473**	**0.7887**	**0.8071**	**0.8240**	**0.8390**

Bold values represent best results.

2.3.6 Supervised model choice

Since model architecture is orthogonal to using a semi-supervised framework, we picked the best-performing model using Dice scores, as can be seen in Table 2. UNet++ performed better on average, exhibited a low variance, often performed the best out of fully-supervised architectures, and almost never underperformed (as shown by the average rank). Detailed results were reported in Tables 3, 4. For these reasons, it was the model we chose to compare to the baseline in the semi-supervised framework.

2.4 Semi-supervised learning (SSL) framework

2.4.1 Cross pseudo supervision (CPS)

As described in Section 2.2, roughly 1% of the collected images were manually annotated and used for training the fully-supervised methods. To take advantage of the potential semantic information contained in the unlabeled images, we used the CPS framework described in Chen et al. (2021). It trained two networks and in addition to using the standard pixel-wise cross-entropy loss on

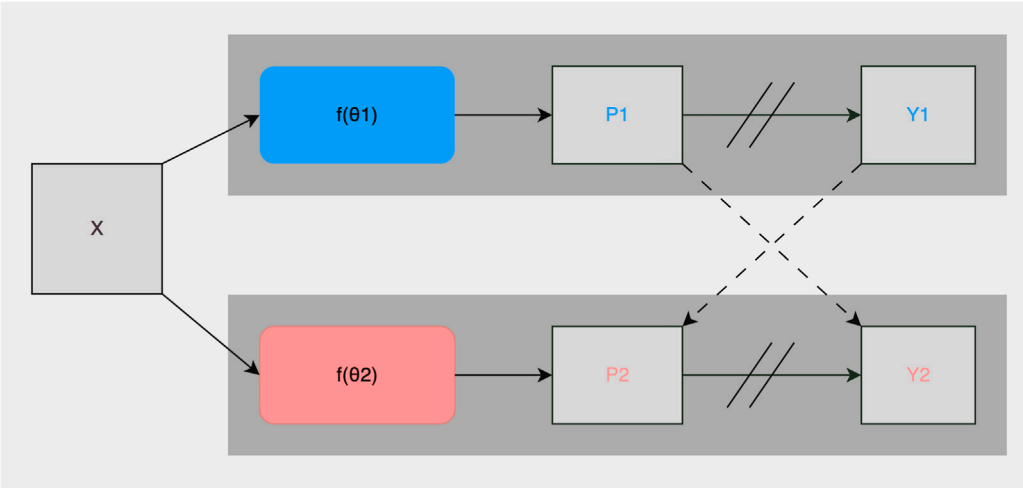

FIGURE 3
Illustration of the CPS training framework. X are inputs with the same data augmentation. θ_1 and θ_2 are two segmentation networks with the same structure and differently initialized weights. P_i is the segmentation confidence map, Y_i is the predicted label map. → means forward operation, −−→ loss supervision, and // on → stopping the gradient.

labeled images, used pseudo-labels generated from the segmentation confidence map of one network to supervise the other as can be seen in Figure 3. Loss for unlabeled images in CPS is defined as follows with D^u denoting unlabeled data, p_i the segmentation confidence map, y_i the predicted label map, ℓ_{ce} the cross-entropy loss, 1, and 2 representing each network:

$$\mathcal{L}_{cps} = \frac{1}{|\mathcal{D}^u|}\sum_{\mathbf{x}\in\mathcal{D}^u}\frac{1}{W\times H}\sum_{i=0}^{W\times H}\left(\ell_{ce}\left(\mathbf{p}_{1i},\mathbf{y}_{2i}\right)+\ell_{ce}\left(\mathbf{p}_{2i},\mathbf{y}_{1i}\right)\right) \quad (2)$$

The total loss is thus defined as follows with $\mathcal{L}_s$ the standard pixel-wise cross-entropy loss:

$$\mathcal{L} = \mathcal{L}_s + \lambda\mathcal{L}_{cps} \quad (3)$$

We trained both ResUNet and UNet++ models with this framework and used the same number of labeled images as in the supervised setting, treating the remaining images in each volume as unlabeled data. In our study, we noticed that models trained with a loss more similar to the evaluation metric performed better. As a consequence, we replaced all of the cross-entropy losses originally used in Chen et al. (2021) with the soft Dice loss 1. We also noticed that learning needed to be driven by the supervised loss during the first epochs. At the beginning of training, models had no prior knowledge of the segmentation task, and thus, could not yield relevant pseudo-labels resulting in frequent collapsing to predict only the background, especially when working with nucleoli. To resolve this, we implemented a linear warm-up to λ, the parameter used to balance the CPS loss with the supervised loss, so that the latter has priority over the former during the early steps of training. We used a value of 1 for λ in all of our experiments and the following function for linear warmup.

$$y = \begin{cases} 0.00001 & \text{if epoch} \leq 5 \\ 0.067x - 0.335 & \text{if } 6 \leq \text{epoch} \leq 19 \\ 1 & \text{otherwise} \end{cases}$$

2.4.2 Integration of CutMix data augmentation

CutMix (Yun et al., 2019) is a popular data augmentation method for training classifiers that shuffles information throughout the training batch, and has recently been used in semi-supervised segmentation tasks. It boosts accuracy by making the models focus on less discriminative parts of objects to segment, in our case we can imagine that some part of a nucleus is replaced with background, then models have to use something other than the nuclear membrane to detect the nucleus, forcing them to learn other features such as texture. In the authors' implementation, when using the CutMix strategy in the CPS loss, the latter is only optionally computed on labeled data. However, in our case, not having labels meant not being able to ensure the CPS batch contained any nucleoli as we did for supervised methods (see Section 2.2). This made the loss unstable as models performed poorly on empty images. To solve this issue, we trained all models with both the supervised and unsupervised CPS loss, and ensured that at least half of the CPS batch contained nucleoli. We tried using CutMix in the fully supervised setting, however, it did not yield any significant improvement. We believe this to be due to the fact that the number of images we trained on was so limited in the fully-supervised setting, CutMix could not add much new information during augmentation.

3 Results and discussion

3.1 Accurate segmentation of nuclei and nucleoli

For comparison with previous results (Machireddy et al., 2021), we first used the same average of 2D dices (Tables 3, 4) while also providing the unbiased 3D Dice metric results (Tables 5, 6). The 3D dice scores are almost always higher than their average counterpart, which is positive as they are more representative of the segmentation

TABLE 5 3D nuclei segmentation dice scores. Numbers 7, 10, 15, and 25 in the second line represent the number of training images for each volume.

	Volume 1				Volume 2				Volume 3			
	7	10	15	25	7	10	15	25	7	10	15	25
UNet++	0.9787	0.9816	0.9845	0.9867	0.9882	0.9913	0.9927	0.9933	0.9639	0.9702	0.9767	0.9849
FracTALResNet	0.9761	0.9825	0.9841	0.9899	0.9765	0.9843	0.9874	0.9934	0.9722	0.976	0.9804	0.9853
CEECNet	0.9749	0.9688	0.985	0.977	0.9852	0.9897	0.993	0.9949	0.9571	0.9592	0.9699	0.9834
Senformer	0.9858	0.9873	0.9897	0.9908	0.9838	0.9872	0.99	0.9927	0.9767	0.9774	0.9829	0.9878
SSL	0.9897	0.9906	0.9898	0.9915	0.9939	0.9931	0.9942	0.9958	0.9687	0.9763	0.9757	0.9848
SSL + CutMix	0.9910	0.9914	0.9918	0.9924	**0.9951**	0.9952	**0.9955**	0.9957	0.9785	0.9793	0.983	0.9849
SSL-UNet++-CutMix	**0.9917**	**0.9922**	**0.9924**	**0.9932**	**0.9951**	**0.9953**	0.9954	**0.9959**	**0.9832**	**0.986**	**0.9878**	**0.9887**

Bold values represent best results.

TABLE 6 3D nucleoli segmentation dice scores. Numbers 7, 10, 15, and 25 in the second line represent the number of training images for each volume.

	Volume 1				Volume 2				Volume 3			
	7	10	15	25	7	10	15	25	7	10	15	25
UNet++	0.9626	0.9647	0.9664	0.9681	0.9073	0.9321	0.9333	0.9362	0.734	0.8115	0.8438	0.869
FracTALResNet	0.9755	0.9657	0.9522	0.9685	0.8814	0.8722	0.8761	0.9371	0.7523	0.8044	0.8365	0.8672
CEECNet	0.9788	0.979	0.979	**0.9806**	0.7852	0.8902	0.8436	0.8899	0.6629	0.7186	0.7927	0.8419
Senformer	0.9396	0.9398	0.9425	0.9453	0.8585	0.8734	0.8962	0.9144	0.7556	0.7778	0.8231	0.8473
SSL	0.9789	0.9797	0.9782	0.9792	0.9109	0.9152	0.9436	0.9513	0.7662	0.806	0.8545	0.8614
SSL + CutMix	**0.9797**	0.9797	**0.98**	0.9801	0.9167	**0.9231**	0.946	0.9525	0.8235	0.8415	0.8558	0.8667
SSL-UNet++-CutMix	0.9794	**0.9798**	0.9793	0.9803	**0.9415**	0.9198	**0.9475**	**0.9536**	**0.8421**	**0.8436**	**0.8626**	**0.8726**

Bold values represent best results.

quality. This is particularly visible for nucleoli (compare Tables 4, 6). As can be observed in Tables 3, 4, all evaluated models were able to accurately segment both nuclei and nucleoli. Despite the introduction of attention enabling models to gain a marginal edge over the baseline introduced in Machireddy et al. (2021), a clear advantage was obtained only when using SSL, especially when paired with CutMix data augmentation. The best results are given by UNet++ with SSL and CutMix, which indicates both model selection and using a semi-supervised framework help improve segmentation performance.

3.2 Benefits of semi-supervised learning

While fully-supervised models could sometimes outperform SSL ones on specific datasets (for example, CEECNet on Volume 1 nucleoli), SSL remained stable over all structures and volumes. It outperformed the baseline for all datasets, most noticeably on Volume 3 nucleoli, with an average gain of 0.11 in Dice, representing a 15.6% performance increase. One of the reasons behind this performance gain is the high heterogeneity in Volume 3 nucleoli, and most models struggled segmenting unseen structures, as can be observed in Figure 4. As the performances of different fully-supervised methods varied highly depending on the volumes (for example, Senformer under-performed in segmenting Volume 1 nucleoli), the SSL methods remained stable.When evaluating our models, we noticed that fully supervised methods performed really well around the images they were trained on (see Figure 5), yielding a near perfect Dice score. However, performances dropped as soon as evaluation images start being dissimilar to the training images, thus forming dips visible in the plot. This is a clear sign of over-fitting that SSL prevented thanks to the regularization added by the CPS loss. This stability and consistency across image volumes allows, in addition to the performance gain, an easier post-processing of the segmented volume by manual inspection and interpretation or algorithmic analysis. These result made us believe semi-supervised frameworks were key in attaining better generalization performance in our sparse annotation setting. Indeed, the Dice score of UNet++ with SSL and Cutmix are better most of the time with only 7 training images than what was achieved previously with 25 images in Machireddy et al. (2021) or with supervised models in this study.

Seeing that the semi-supervised framework yielded way more consistent results, we looked into whether segmentation networks trained under this framework would be able to generalize across samples. To test this, we trained a model on two of our three volumes and evaluated on the third. We noticed that the performance

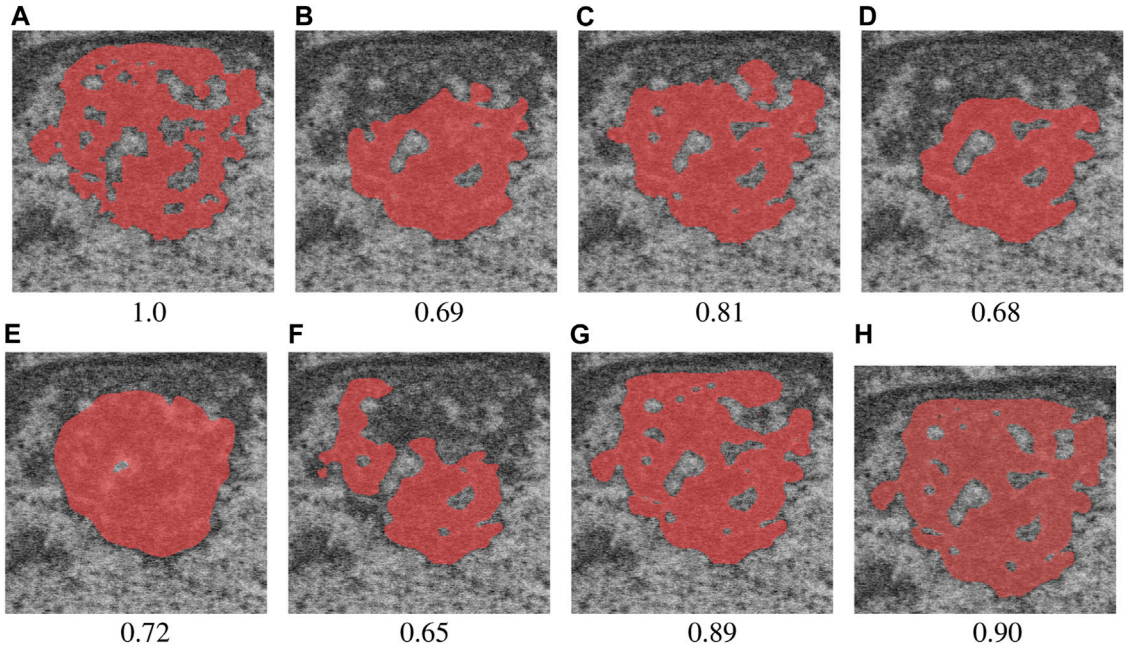

FIGURE 4
Qualitative results with Dice score for a difficult nucleolus in Volume 3, from **(A)** ground truth, **(B)** UNet++, **(C)** FracTALResNet, **(D)** CEECNet, **(E)** Senformer, **(F)** SSL-ResUNet, **(G)** SSL-ResUNet-CutMix, and **(H)** SSL-UNet++-CutMix.

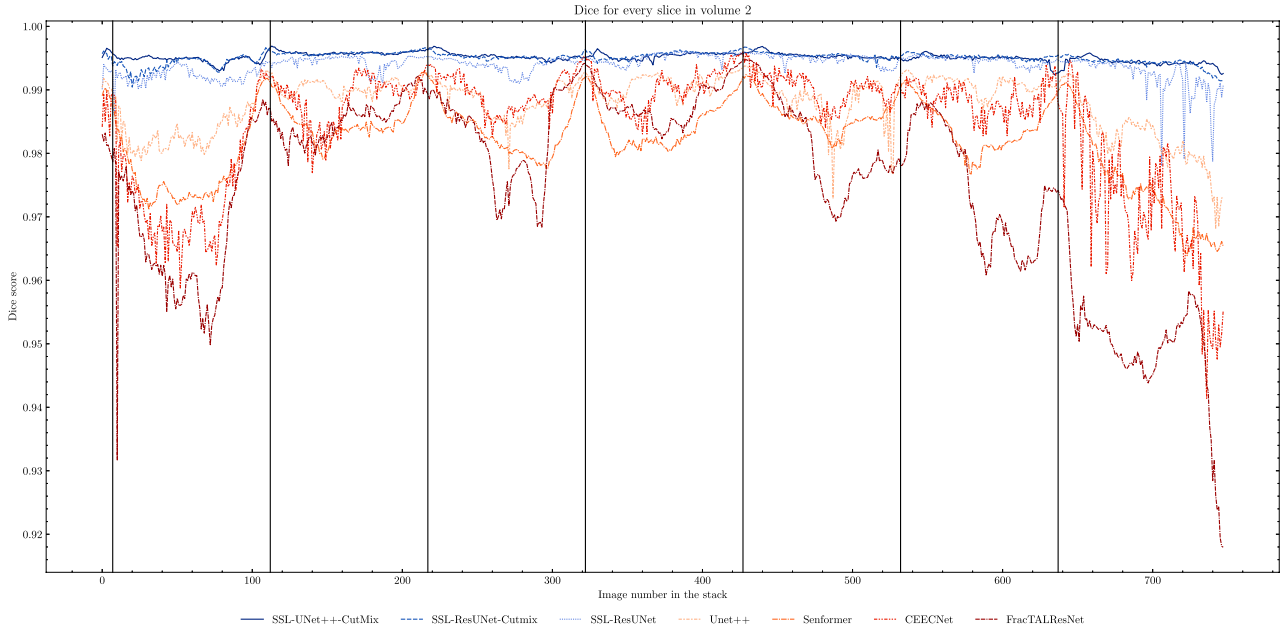

FIGURE 5
Comparison of Dice scores for nuclei segmentation along all 757 slices in Volume 2 with 7 training images. Training slices are marked with vertical black lines. We can clearly observe the seven peaks in performance and drops in between for fully-supervised methods (beige to brown) as opposed to the stability of the SSL models (blue).

dropped significantly (with dice scores between 0.4 and 0.6) depending on the volume. We hypothesize that two volumes worth of data is not enough to generalize across samples and leave to future work exploration of ways to deal with this issue such as expanding the dataset, data normalization, knowledge distillation or style transfer.

4 Conclusion

In this work, we investigated the segmentation of nuclei and nucleoli in vEM images of cancer cells. We studied the performances of several leading deep-learning models and assessed the relative performance gains of each method. We provided insight as to why semi-supervised methods were able to yield more robust results and managed to improve on previous work both in terms of reducing the amount of data needed and segmentation performances, with an improved Dice on all Volumes. We believe the key components that make these improvements possible in our pipeline are the combination of exploiting unlabeled data information (SSL framework), carefully picking a segmentation model, effective data augmentation methods (CutMix), efficient losses, and keeping dataset specificities in mind (e.g., adapting the batch selection and CPS framework to nucleoli sparsity (see Sections 2.2, 2.4.1, 2.4.2). We made the experiment code available at[3] and the complete manual annotations for the data have been provided through the HTAN data portal. We believe that semi-supervised methods are a key component in segmentation with sparse annotations as they proved to be superior in both quantitative and qualitative evaluations.

Data availability statement

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by the Oregon Health and Science University Institutional Review Board. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

LP: Formal Analysis, Investigation, Methodology, Software, Visualization, Writing–original draft, Writing–review and editing. GT: Investigation, Supervision, Writing–review and editing, Formal Analysis, Methodology, Software. WB: Software, Writing–review and editing. JR: Conceptualization, Data curation, Investigation, Project administration, Supervision, Visualization, Writing–review and editing. XS: Conceptualization, Formal Analysis, Investigation, Methodology, Supervision, Writing–original draft, Writing–review and editing. JG: Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing–review and editing.

3 https://github.com/LucasPagano/Segmentation3DEM

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This manuscript was supported by the NCI Human Tumor Atlas Network (HTAN) Omic and Multidimensional Spatial (OMS) Atlas Center Grant (5U2CCA233280), Prospect Creek Foundation, the Brenden-Colson Center for Pancreatic Care, the NCI Cancer Systems Biology Measuring, Modeling, and Controlling Heterogeneity (M2CH) Center Grant (5U54CA209988), the OHSU Knight Cancer Institute NCI Cancer Center Support Grant (P30CA069533). This work is supported by the Cancer Early Detection Advanced Research (CEDAR) Center at the Knight Cancer Institute at OHSU and the OCS.

Acknowledgments

FIB-SEM data included in this manuscript were generated at the Multiscale Microscopy Core (MMC), an OHSU University Shared Resource, with technical support from the OHSU Center for Spatial Systems Biomedicine (OCSSB). The authors acknowledge the Knight Cancer Institute's Precision Oncology SMMART Clinical Trials Program for the resources, samples, and data that supported this study. Specimen acquisition support from the SMMART clinical coordination team was invaluable. This study was approved by the Oregon Health and Science University Institutional Review Board (IRB#16113). Participant eligibility was determined by the enrolling physician and informed consent was obtained prior to all study-protocol-related procedures.

Conflict of interest

JG has licensed technologies to Abbott Diagnostics; has ownership positions in Convergent Genomics, Health Technology Innovations, Zorro Bio and PDX Pharmaceuticals; serves as a paid consultant to New Leaf Ventures; has received research support from Thermo Fisher Scientific (formerly FEI), Zeiss, Miltenyi Biotech, Quantitative Imaging, Health Technology Innovations and Micron Technologies; and owns stock in Abbott Diagnostics, AbbVie, Alphabet, Amazon, Amgen, Apple, General Electric, Gilead, Intel, Microsoft, Nvidia, and Zimmer Biomet.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

References

Baghban, R., Roshangar, L., Jahanban-Esfahlan, R., Seidi, K., Ebrahimi-Kalan, A., Jaymand, M., et al. (2020). Tumor microenvironment complexity and therapeutic implications at a glance. *Cell. Commun. Signal* 18, 59. doi:10.1186/s12964-020-0530-4

Bousselham, W., Thibault, G., Pagano, L., Machireddy, A., Gray, J., Chang, Y. H., et al. (2022). "Efficient self-ensemble for semantic segmentation," in 33rd British Machine Vision Conference 2022, BMVC 2022, London, United Kingdom, November 21–24, 2022. Available at: https://bmvc2022.mpi-inf.mpg.de/0892.pdf.

Bushby, A. J., Young, R. D., Pinali, C., Knupp, C., and Quantock, A. J. (2011). Imaging three-dimensional tissue architectures by focused ion beam scanning electron microscopy. doi:10.1038/nprot.2011.332

Chen, X., Yuan, Y., Zeng, G., and Wang, J. (2021). "Semi-supervised semantic segmentation with cross pseudo supervision," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2021), Nashville, TN, Virtual/Online due to COVID-19.

Diakogiannis, F. I., Waldner, F., and Caccetta, P. (2021). Looking for change? roll the dice and demand attention. *Remote Sens.* 2021, 3707. doi:10.3390/RS13183707

Giannuzzi, L., and Stevie, F. (2005). *Introduction to focused ion beams: instrumentation, theory, techniques and practice*. Berlin: Springer. doi:10.1007/0-387-23313-X_2

Glorot, X., and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. *J. Mach. Learn. Res. - Proc. Track* 9, 249–256.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 770–778. doi:10.1109/CVPR.2016.90

Hirata, E., and Sahai, E. (2017). Tumor microenvironment and differential responses to Therapy. *Cold Spring Harb. Perspect. Med.* 7, a026781. doi:10.1101/cshperspect.a026781

Jain, J., Singh, A., Orlov, N., Huang, Z., Li, J., Walton, S., et al. (2021). Semask: semantically masked transformers for semantic segmentation

Karabağ, C., Jones, M. L., Peddie, C. J., Weston, A. E., Collinson, L. M., and Reyes-Aldasoro, C. C. (2020). Semantic segmentation of hela cells: an objective comparison between one traditional algorithm and four deep-learning architectures. *PLoS ONE* 15, e0230605. doi:10.1371/JOURNAL.PONE.0230605

Karnovsky, M. (1964). A formaldehyde-glutaraldehyde fixative of high osmolality for use in electron microscopy. *J. Cell. Biol.* 27.

Lindström, M. S., Jurada, D., Bursac, S., Orsolic, I., Bartek, J., and Volarevic, S. (2018). Nucleolus as an emerging hub in maintenance of genome stability and cancer pathogenesis. *Oncogene* 37, 2351–2366. doi:10.1038/s41388-017-0121-z

Machireddy, A., Thibault, G., Loftis, K. G., Stoltz, K., Bueno, C. E., Smith, H. R., et al. (2021). Robust segmentation of cellular ultrastructure on sparsely labeled 3d electron microscopy images using deep learning. *bioRxiv*. doi:10.1101/2021.05.27.446019

Nunney, L., Maley, C. C., Breen, M., Hochberg, M. E., and Schiffman, J. D. (2015). Peto's paradox and the promise of comparative oncology. *Philosophical Trans. R. Soc. B Biol. Sci.* 370, 20140177. doi:10.1098/rstb.2014.0177

Perez, A. J., Seyedhosseini, M., Deerinck, T. J., Bushong, E. A., Panda, S., Tasdizen, T., et al. (2014). A workflow for the automatic segmentation of organelles in electron microscopy image stacks. *Front. Neuroanat.* 8, 126. doi:10.3389/fnana.2014.00126

Riesterer, J. L., López, C. S., Stempinski, E. S., Williams, M., Loftis, K., Stoltz, K., et al. (2019). A workflow for visualizing human cancer biopsies using large-format electron microscopy. *bioRxiv*. doi:10.1101/675371

Shi, J., Kantoff, P. W., Wooster, R., and Farokhzad, O. C. (2017). Cancer nanomedicine: progress, challenges and opportunities. *Nat. Rev. Cancer* 17, 20–37. doi:10.1038/nrc.2016.108

Siddique, N., Paheding, S., Elkin, C. P., and Devabhaktuni, V. (2020). U-net and its variants for medical image segmentation: a review of theory and applications. *IEEE Access* 9, 82031–82057. doi:10.1109/access.2021.3086020

Su, R., Zhang, D., Liu, J., and Cheng, C. (2021). Msu-net: multi-scale u-net for 2d medical image segmentation. *Front. Genet.* 12, 639930. doi:10.3389/fgene.2021.639930

Taha, A. A., and Hanbury, A. (2015). Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. *BMC Med. Imaging* 15, 29. doi:10.1186/s12880-015-0068-x

Tanaka, H. Y., and Kano, M. R. (2018). Stromal barriers to nanomedicine penetration in the pancreatic tumor microenvironment. *Cancer Sci.* 109, 2085––2092. doi:10.1111/cas.13630

Thevenaz, P., Ruttimann, U., and Unser, M. (1998). A pyramid approach to subpixel registration based on intensity. *IEEE Trans. Image Process.* 7, 27–41. doi:10.1109/83.650848

Yun, S., Han, D., Chun, S., Oh, S. J., Choe, J., and Yoo, Y. (2019). "Cutmix: regularization strategy to train strong classifiers with localizable features," in Proceedings of the IEEE International Conference on Computer Vision, China, 2019-October (IEEE), 6022. doi:10.1109/ICCV.2019.00612

Zhang, F., Breger, A., Cho, K. I. K., Ning, L., Westin, C. F., O'Donnell, L. J., et al. (2021). Deep learning based segmentation of brain tissue from diffusion mri. *NeuroImage* 233, 117934. doi:10.1016/J.NEUROIMAGE.2021.117934

Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., and Liang, J. (2018). Unet++: a nested u-net architecture for medical image segmentation. *Corr. abs/1807*, 10165. doi:10.1007/978-3-030-00889-5_1

Zink, D., Fischer, A. H., and Nickerson, J. A. (2004). Nuclear structure in cancer cells. *Nat. Rev. Cancer* 4, 677–687. doi:10.1038/NRC1430