Deep Learning Frameworks for Electron Microscopy Image Analysis with Sparse Labels

Lucas Pagano

Presented in partial fulfillment

of the requirements for the degree of

Master of Science in

Biomedical Engineering

to the **School of Medicine** at

Oregon Health & Science University.

Biomedical Engineering

School of Medicine

Oregon Health & Science University

December, 2024

Biomedical Engineering
School of Medicine
Oregon Health & Science University

CERTIFICATE OF APPROVAL

This is to certify that the MS thesis of
Lucas Pagano
has been approved.

Xubo Song
Advisor

Kyle Ellrott
Chair

Jessica L. Riesterer
Committee Member

Young Hwan Chang
Committee Member

TABLE OF CONTENTS

ABSTRACT

This thesis presents a suite of deep learning frameworks designed to address key challenges in the analysis of Electron Microscopy (EM) images of cancer cells, which are vital for understanding cancer progression and therapy resistance. The research focuses on three primary domains: (1) Image Quality Assessment and Denoising, where automated metrics and deep learning models validate faster, safer EM sample preparation protocols without compromising image quality; (2) Semi-supervised Semantic Segmentation, leveraging limited manual annotations to train models for accurate 3D segmentation of nuclei and nucleoli, essential for cancer diagnostics; and (3) Object Detection and Unsupervised Segmentation, integrating advanced detection and segmentation methods to analyze complex organelles such as mitochondria and endosomes at nanoscale resolution. The findings demonstrate the efficacy of these approaches in reducing annotation bottlenecks and enhancing the robustness of EM image analysis. By automating key aspects of image processing, this work contributes significantly to accelerating cancer research and supporting clinical applications.

Acknowledgements

I would like to express my deepest gratitude to my advisor and mentor Xubo Song, for her invaluable guidance, unwavering support, and insightful advice throughout the course of this research. Her mentorship has been instrumental in shaping this thesis and my development as a researcher.

I would like to thank Monica Hinds for her support through out and especially over the last few months in guiding and helping me through the process of writing and defending this thesis. I appreciate everything she has done for me over the years to ensure I could continue my research.

I am profoundly thankful to Young Hwan Chang for his expertise in image-processing algorithms within the context of medical image processing. His contributions were a tremendous source of knowledge and greatly enriched this work. I am equally grateful to Guillaume Thibault for his hands-on assistance with technical challenges and his meticulous support, both of which were critical to the completion of this thesis.

I extend my sincere appreciation to Jessica L. Riesterer for her help in elucidating the image acquisition pipeline and the underlying biological principles, which provided essential context and depth to my research. I also thank Cecilia Bueno, Hannah Smith and Sanjay Srikanth with their help generating the ground truth data.

I would also like to thank Kyle Ellrott for his valuable insights, thoughtful feedback, and steadfast support throughout the entire duration of this project.

Finally, I would like to acknowledge George Thomas for his role as a member of my initial advisory committee. His expertise and guidance, particularly in the biological aspects of this work, were greatly appreciated and laid a strong foundation for this research.

To all my committee members, Xubo Song, Young Hwan Chang, Jessica L. Riesterer, and Kyle Ellrott, I am deeply grateful for your collective contributions and commitment to supporting my academic journey. Your advice and encouragement have been pivotal in the realization of this thesis.

List of Figures

List of Tables

List of Abbreviations

Symbols

2D Two-Dimensional. 10

3D Three-Dimensional. 3, 10, 13, 20

B

BRISQUE Blind/Referenceless Image Spatial Quality Evaluator. 5

D

DCIS Ductal Carninoma In Situ. 1, 2

DISTS Deep Image Structure and Texture Similarity. 6

E

EM Electron Microscopy. iii, 1, 2, 11, 20–22, 28–30

F

FIB-SEM Focused Ion Beam Scanning Electron Microscopy. 4

G

GMSD Gradient Magnitude Similarity Deviation. 6

H

HTAN Human Tumor Atlas Network. 11, 18

I

IoU Intersection over Union. 20, 24, 27

IQ Image Quality. 5, 6, 8

IQA Image Quality Assessment. 3–5, 31

M

mAP mean Average Precison. 20, 24, 27–29

MSSIM Multi-scale Structural Similarity Index Measure. 5

P

PDAC Pancreatic ductal adenocarcinoma. 1

PSNR Peak Signal-to-Noise Ratio. 4, 5

S

SBF-SEM Serial Block Face Scanning Electron Microscopy. 4

SR-SIM Spectral Residual Based Similarity Index Measure. 6, 8

SSIM Structural Similarity Index Measure. 4, 5, 8

V

vEM volume Electron Microscopy. 4, 10

1 Introduction

1.1 Problem statement

Cancer is known for its high morphological and behavioral heterogeneity amongst subtypes, patients and grades [1, 2, 3], and understanding this heterogeneity is key for prognosis and therapy[4, 5, 6, 7]. Nuclear grade has been shown to be tied to tumor biology and correlated to biological prognostic variables[8, 9]. Furthermore, aberrant nuclei and nucleoli are both commonly used cancer markers[10] and the study of structural nucleoli changes has recently emerged as a promising therapeutic approach for cancer treatment[11]. Mitochondrial dysfunction is often viewed as a cause tumorogenesis [12] and endosomes are responsible for cell autophagy and cell signaling, both processes that can favor cancer development when interrupted [13]. Thus, accurate detection and quantization of these organelles will enable a deeper understanding of the underlying mechanisms taking place during cancer development.

To further our comprehension of cancer, several imaging techniques have been developed. As the cellular components to be imaged often exist in the nanometer space, researchers have been using EM to get nanometer resolution views of cellular interactions and couple them to other imaging techniques to get multi-scale representations, with EM already being implemented as a pathology tool, particularly in the renal community[14]. EM was also used for morphological differentiation of cancer stem cells in Pancreatic ductal adenocarcinoma (PDAC)[15] and characterization of structural features for breast cancer stratification[16].

However, while acquiring EM images is now a routine task, the limiting step in the analysis of collected images is the extraction of meaningful features which is currently done by experts through hand annotating. While yielding effective results, it is a time-consuming task, making it unsuitable for biological applications and decisions where time is a critical factor. Indeed, hand-annotating a single sample can take months. Moreover, even without considering the annotation process, a single sample surface typically contains between 500 and 1000 images per Two-Dimensional (2D) tiled montage, of 6000 per 4000 pixels each, where up to 1000 organelles can live in one montage tile [17]. This makes computing spatial and morphological statistics impractical for a human being, and they will have a general impression of a sample rather than precise metrics. On the other hand, an algorithm can process images much quicker, thus motivating the use of algorithmic solutions not only for annotation but also analysis of collected images. As such, to be able to fully leverage the possibilities offered by both deep learning models and high-resolution imaging, developing automated and robust models to replace humans in both the delineating and analysis task is critical[18]. Typically, these models are chained together into an image processing pipeline. A generic example of such a pipeline can be seen in Figure 1.

In Figure 2, we compare images taken from normal tissue with triple negative breast cancer sample and Ductal Carninoma In Situ (DCIS) sample that highlight structural features differentiating between normal and cancer tissues and cancer grades. In both 2a and 2b, nuclei contours appear smoother, with no to small invaginations, in clear opposition with 2c, 2e, 2f where irregular contours and deep invaginations can be observed. Furthermore, the nucleoli

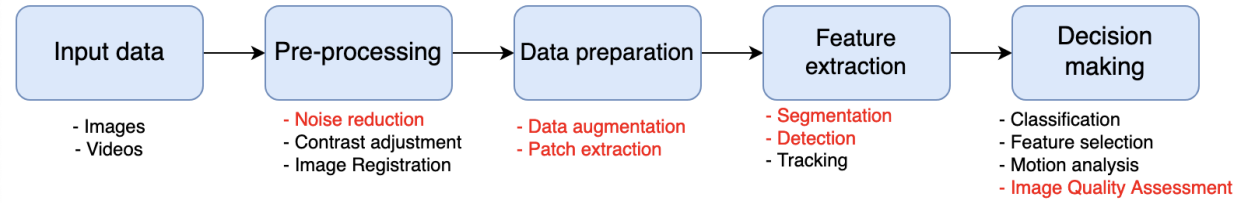
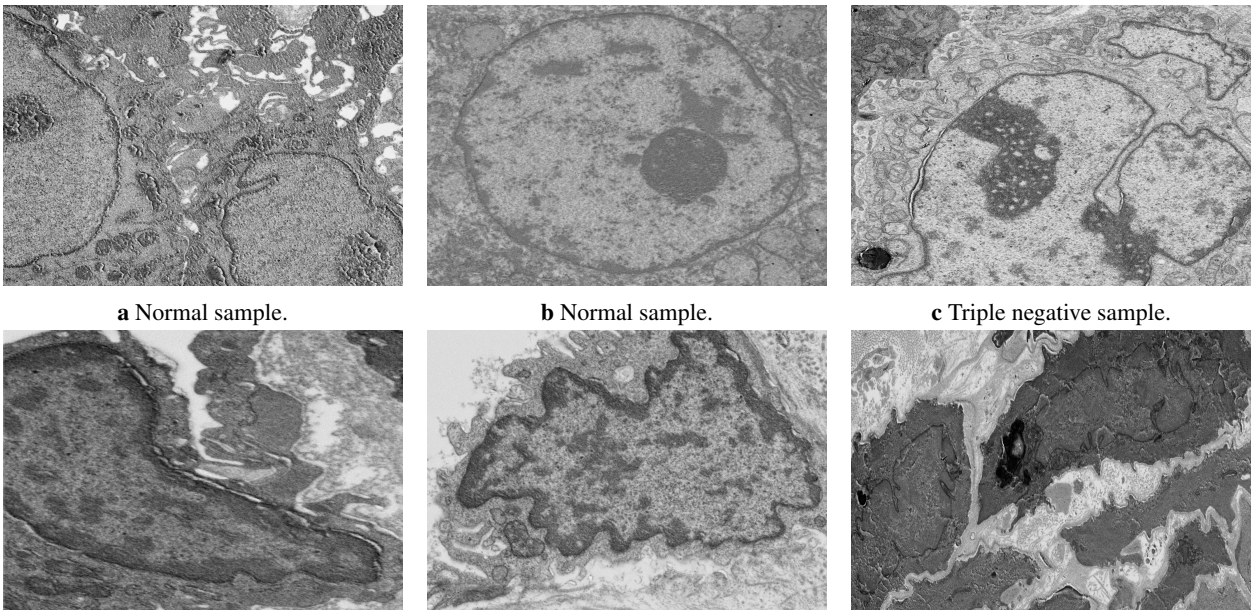


Figure 1: Generic image processing pipeline for medical images. The red steps are dealt with in this thesis.



a Normal sample. **b** Normal sample. **c** Triple negative sample.

d DCIS sample. **e** DCIS sample. **f** DCIS sample.

Figure 2: EM Images displaying actionable features for cancer detection and stratification. Resolution is $4\,\mathrm{nm}$ per pixel, respective image pixel widths heights are 1300×1214 (a) 1108×1044 (b) 1414×1500 (c) 572×716 (d) 884×668 (e) 2576×1106 (f).

in images 2a and 2b are smooth and not fenestrated, in contrast to the high fenestration observed in 2c. As DCIS is an early stage of cancer, cells found in the DCIS sample can appear normal-like as can be seen in **Fig** 2d and be indicative of a low-grade DCIS. They can also appear slightly mutated and be indicative of an intermediate-grade DCIS as in 2e or appear heavily mutated with deep invaginations and be indicative of a high-grade DCIS as in 2f. This variability highlights how statistics (e.g., counting of the low intermediate and high-grade nuclei) over the entirety of the collected images can be beneficial as opposed to computing them only on labeled images. Specific features also necessitate going through several images from different samples to be actionable. For instance, nuclei size variability in samples of normal tissue is way lower than in samples of cancer tissue. It is important to note that some useful features for cancer detection and stratification that have not yet been identified could reside in EM images. Thanks to having meta-data about the samples (e.g., results of an immunochemistry test), it is possible to test if EM images can be used to predict such meta-data and thus if this information is inside the EM images, which cannot necessarily be done by humans, even experts.

1.2 Contributions and Overview

1.2.1 Image Quality Assessment and Denoising

In chapter 2, we introduce a deep learning-based approach to Image Quality Assessment (IQA) and denoising of Electron Microscopy images. We used this approach to evaluate if images produced by protocols with faster process times, less exposure of the operator to hazardous and toxic chemicals and improved reproducibility of the specimens' heavy metal staining retain the same quality that those of a well-known benchmark protocol. The experimental results showed that faster protocols can indeed be used to collect images of the same quality than those collected with the bench protocol for a variety of samples and protocol needs.

1.2.2 Semi-supervised segmentation

In chapter 3, we introduce a semi-supervised framework to improve on previous results obtained with the ResUNet architecture for semantic segmentation of nuclei and nucleoli in Three-Dimensional (3D) stacks of Electron Microscopy images of cancer cells. We benchmark several state of the art fully-supervised models such as UNet++, FracTALResNet, SenFormer, CEECNet and the semi-supervised Cross Pseudo Supervision framework to draw conclusions on the relative gains of using more complex models, semi-supervised learning as well as next steps for the mitigation of the manual segmentation bottleneck. We gained an insight as to why semi-supervised models are able to gain as much as 15.6% relative performance increase over fully-supervised models and establish guidelines for future work on high variability images such as those of tumor cells.

1.2.3 Object detection and unsupervised segmentation

In chapter 4, we introduce a framework for unsupervised segmentation of nuclei, nucleoli, mitochondria and endosomes, using fully-supervised detection. As discussed in section 1.1, while acquiring Electron Microscopy images is now a routine task, their annotation and analysis represents the bottleneck of the image processing pipeline. It is even more true for semantic segmentation annotations, where every pixel in the image needs to be annotated, as opposed to object detection, where an object is only associated with 4 coordinates corresponding to its bounding box. We made use of the ease of annotation of objects in the object detection paradigm to collect a collection of 42 samples with varied cancer subtypes, tissue types and cancer grades for breast cancer with annotations for the aforementioned organelles. This large volume of annotated data enabled us to train robust detection models and get good detection results with especially strong results for the detection of nuclei. In particular, we report a mAP of 0.683, which is comparable to results on the natural images COCO detection dataset with full labels [19]. Moreover, our reported AP50 for all organelles outperforms what is reported in state of the art methods such as MitoNet [20] on most of their datasets, especially the cancer images (HeLa). We improved organelle detection AP50 by at least 0.20 mAP which is a considerable improvement. We then leverage these object detection models with the Segment Anything Model (SAM) [21] to obtain segmentation masks for nuclei, nucleoli, mitochondria and endosomes, gaining access to finer features.

2 Deep learning-based Image Quality Assessment and Denoising

2.1 Abstract

New developments in electron microscopy technology, improved efficiency of detectors, and artificial intelligence applications for data analysis over the past decade have increased the use of volume Electron Microscopy (vEM) in the life sciences field. Moreover, sample preparation methods are continuously being modified by investigators to improve final sample quality, increase electron density, combine imaging technologies, and minimize the introduction of artifacts into specimens under study. There are a variety of conventional bench protocols that a researcher can utilize, though most of these protocols require several days. In this work, we describe the utilization of an automated specimen processor, the mPrep™ ASP-2000™, to pre-pare samples for vEM that are compatible with Focused Ion Beam Scanning Electron Microscopy (FIB-SEM), Serial Block Face Scanning Electron Microscopy (SBF-SEM), and array tomography (AT). The protocols assessed here aimed for methods that are completed in a much shorter period of time while minimizing the exposure of the operator to hazardous and toxic chemicals and improving the reproducibility of the specimens' heavy metal staining, all without compromising the quality of the data acquired using backscattered electrons during SEM imaging. As a control, we have included a widely used sample bench protocol and have utilized it as a comparator for image quality analysis, both qualitatively and using image quality analysis metrics. [1]

2.2 Introduction

The quality of images acquired through vEM is a critical factor in the success of downstream analyses, particularly in high-resolution structural studies of biological specimens. The field has increasingly turned to automated approaches to streamline sample preparation while ensuring reproducible outcomes, as demonstrated by techniques such as en-bloc staining and heavy metal impregnation in electron microscopy [22]. However, maintaining consistent image quality across varied samples and processing conditions remains a significant challenge.

Image Quality Assessment (IQA) methodologies are integral to evaluating the effectiveness of these protocols, minimizing human bias, and guiding optimization efforts. Traditional IQA metrics like Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) offer insights into image fidelity but are often insufficient to capture the complexity of vEM datasets, necessitating the use of advanced computational tools, including deep learning approaches [23].

Recent advancements, such as the incorporation of automated specimen processors like the ASP-2000, further underline the role of automation in addressing variability. This system enables precision in sample handling and staining while

[1] The work in this chapter was published in Erin S. Stempinski, Lucas Pagano, Jessica L. Riesterer, Steven K. Adamou, Guillaume Thibault, Xubo Song, Young Hwan Chang, and Claudia S. López. Chapter 1 - automated large volume sample preparation for vem. In Volume Electron Microscopy, volume 177 of Methods in Cell Biology, pages 1–32. Academic Press, 2023.

integrating with IQA pipelines for real-time quality monitoring and protocol refinement [24]. By combining such automated platforms with robust IQA frameworks, researchers can more effectively ensure consistency, enabling broader applications across diverse specimen types, such as neural tissue or cancer models [25].

2.3 Materials and methods

In an effort to quantify image quality with respect to protocol used, but without introducing human bias via sample visible inspection, computer-aided models to evaluate several image quality metrics were implemented. We tried evaluating the metrics both on entire stitched tilesets and on individual tiles. As results were very similar, we only show the latter here. It is important to note that evaluating the collected images is difficult; as they are not from the same sample, they do not image the same region of interest, and thus their content is dissimilar. Since a pixel-to-pixel comparison of images is impossible because they do not represent the exact same area of interest, we partially annotated resin in each of them, as can be seen in Figure 3.

As resin is an identical external material added to each sample, this enables noise evaluation by computing the signal's standard deviation in the annotated parts without having to consider signal disparities. Thus, lower values of standard deviation will mean less noise and better Image Quality (IQ). We also use Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [26], which allows the evaluation of IQ without a reference image (ground truth), and yields a score between 0 and 100, where lower values are indicative of better IQ. This method is based on scene statistics of locally normalized luminance coefficients to quantify possible losses of "naturalness" in the image due to the presence of distortions. BRISQUE has been shown to be highly competitive and computationally more efficient than other No-Reference IQ metrics on medical images [27].

Most IQ metrics use a reference image to compare with the degraded image to analyze, but in our case these references do not exist. However, we were able to train a deep learning-based denoising model called Noise2Void to remove the noise from our images without requiring any ground truth [28]. We then used the denoised images as references to their noisy counterparts and compare them using the IQ metrics described below:

- Peak Signal-to-Noise Ratio (PSNR): is based on Mean Standard Error and is a point-wise IQ metric. Typical values for PSNR range between 30 and 50 dB, where higher is better. It is frequently used in case studies and benchmarks as a weak evaluator baseline, but it is known to give results far from human perception [23].

- Structural Similarity Index Measure (SSIM): is based on luminance, contrast and structure, while introducing the concept of inter-dependency between spatially close pixels by being computed on various windows in the image [23, 29]. Values range from 0 to 1, with 1 being perfect structural similarity.

- Multi-scale Structural Similarity Index Measure (MSSIM) is a multi-scale version of SSIM which introduces an image synthesis approach that automatically determines the relative importance of each scale [30]. Like SSIM, values range from 0 to 1, with 1 indicating best quality.

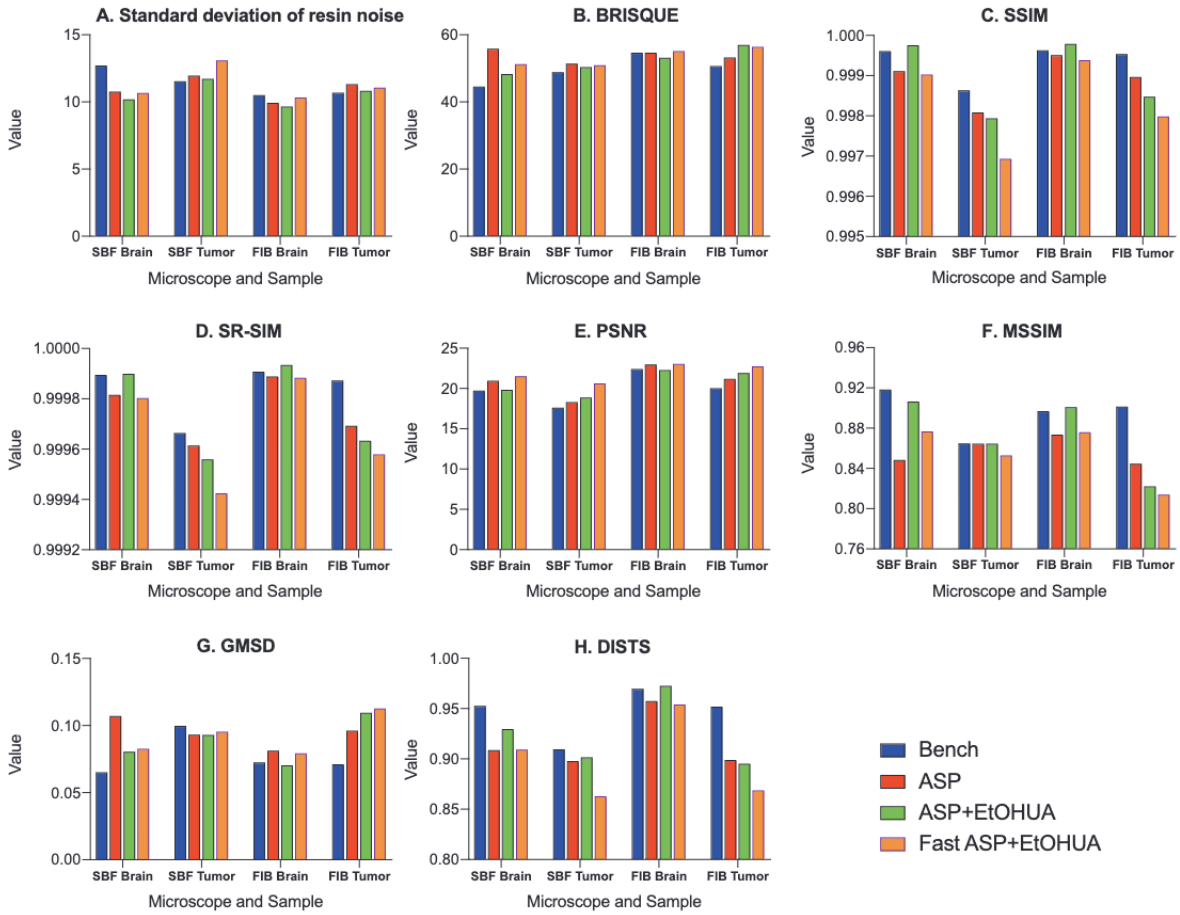


Figure 3: Examples of hand-annotated resin for noise measurements. Areas annotated are marked by an asterisk and blue coloration. (A) Annotated resin in bench-processed brain imaged with the SBF-SEM. Scale bar = $40\,\mu\mathrm{m}$. (B) Annotated resin in ASP-2000 mouse tumor imaged with the FIB-SEM. Scale bar = $20\,\mu\mathrm{m}$.

- Spectral Residual Based Similarity Index Measure (SR-SIM) compares the spectral residual saliency maps of the images. SR-SIM is designed on the hypothesis that an image's saliency map is closely related to its perceived quality [31]. Values range from 0 to 1, with 1 being the best possible score.
- Gradient Magnitude Similarity Deviation (GMSD) valuates the difference between the images' gradients. GMSD is based on the pixel-wise gradient magnitude similarity (GMS) and a novel pooling strategy: instead of using the average as is usually done (and works poorly because it ignores the difference in quality degradation relative to each area), the final score is given by the standard deviation of the Gradient Magnitude Similarity map, which is the range of distortions severities between images: the higher the GMSD score, the larger the distortion range, thus the lower the IQ [32]. Values range from 0 to 1, where lower is better.
- Deep Image Structure and Texture Similarity (DISTS) makes use of neural networks to assess IQ. It has been shown to give a closer evaluation of human quality perception than other previously described IQ metrics. In

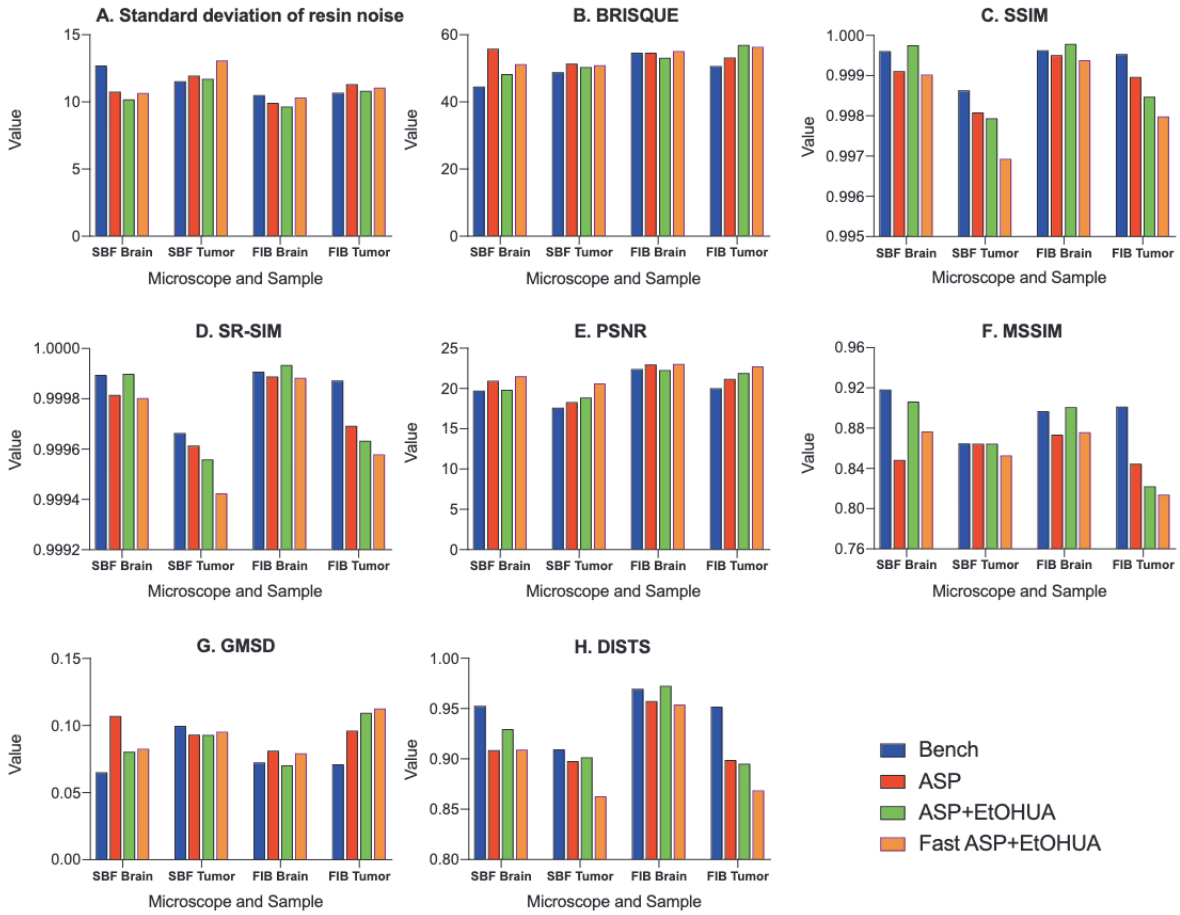
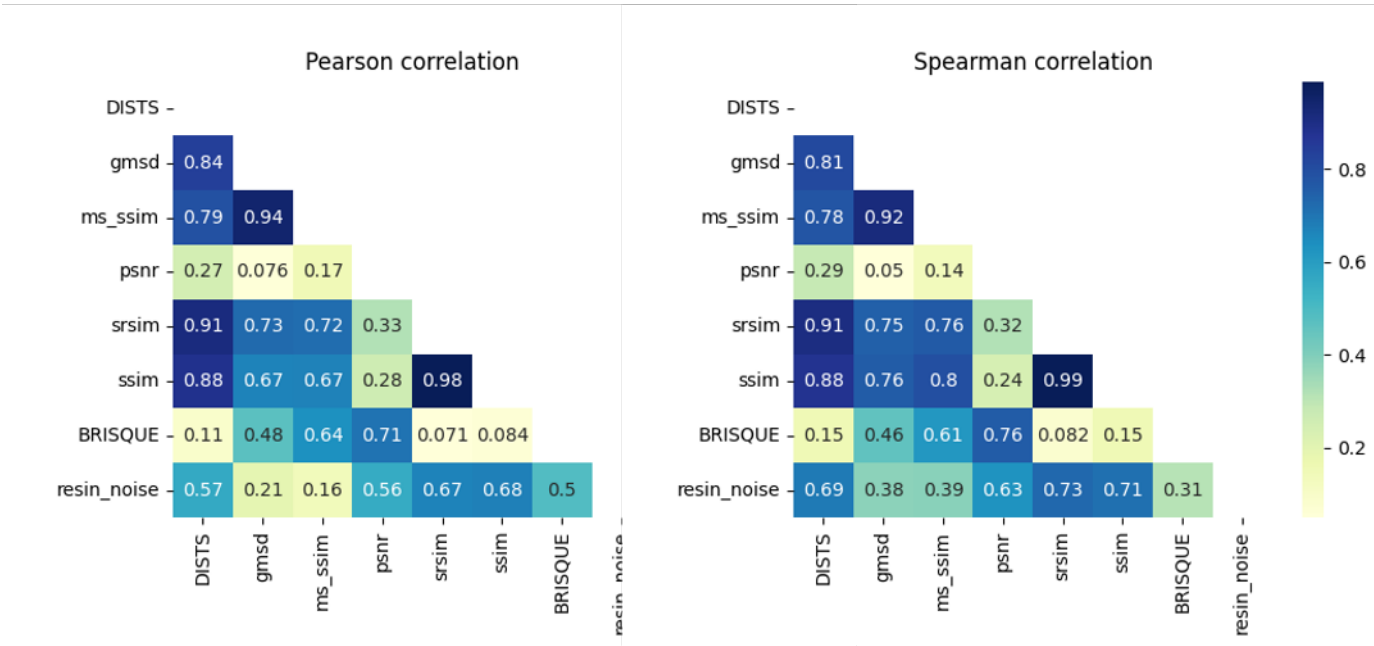


Figure 4: Image quality metrics for brain and tumor samples imaged with the SBF-SEM (SBF) or FIB-SEM (FIB) imaging platforms processed with the bench, ASP-2000 (ASP) ASP-2000 with ethanolic UA (ASP+EtOHUA) or fast ASP-2000 with ethanolic UA (fast ASP+EtOHUA) protocols. The following metrics were evaluated and compared to the standard deviation of noise in resin (A), (B) Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE), (C) Structural Similarity Index Measure (SSIM), (D) Spectral Residual Based Similarity Index Measure (SR-SIM), (E) Peak Signal-to-Noise Ratio (PSNR), (F) Multi-scale Structural Similarity Index Measure (MSSIM), (G) Gradient Magnitude Similarity Deviation (GMSD), and (H) Deep Image Structure and Texture Similarity (DISTS).

particular, it is less sensitive to point-by-point deviations between the images [33]. Values range from 0 to 1, where 1 is a perfect score.

2.4 Results and conclusion

For our image analysis, we used noise measurements on blank resin as a way to understand image quality, as noise may disrupt the ability of machine learning algorithms that are frequently utilized in our community to segment cellular features (Figure 4A). We additionally compared our noise measurements to other selected image quality metrics.



a Pearson correlation matrix for IQ metrics.

b Spearman correlation matrix for IQ metrics.

Figure 5: Pearson (A) and Spearman (B) correlation matrices between image quality metrics. The following metrics were compared: the standard deviation of noise in resin, Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE), Structural Similarity Index Measure (SSIM), Spectral Residual Based Similarity Index Measure (SR-SIM), Peak Signal-to-Noise Ratio (PSNR), Multi-scale Structural Similarity Index Measure (MSSIM), Gradient Magnitude Similarity Deviation (GMSD), and Deep Image Structure and Texture Similarity (DISTS).

As can be observed in Figure 4 our results indicate that, for the metrics we selected, patterns emerged dependent on sample type. For brain samples in both the SBF-SEM and FIB-SEM images analyzed, the ASP-2000 protocol with ethanolic UA resulted in similar and occasionally better image quality scores when compared to the other protocols tested. For tumor samples, the bench protocol tended to have the best image quality scores, followed by the ASP-2000 protocol, the ASP-2000 protocol with ethanolic UA, and finally the fast ASP-2000 protocol.

The metric that was the most straightforward and reliable in our analysis was the standard deviation in the hand-annotated resin areas, as we are sure it only considers noise and not signal (Figure 4A). SSIM and SR-SIM have very small variations, which could mean they are not suited for the task, but they are correlated with most other metrics (Figure 4C and D).

For further analysis, we created Pearson and Spearman correlation matrices between the metrics analyzed (Figure 5). The standard deviation in resin is the closest that we have to a ground truth evaluation of the noise. This result strongly correlated to DISTS, SRSIM and SSIM, indicating that these metrics are good evaluators of the noise in our images. On the other hand, BRISQUE and PSNR exhibit a strong correlation but are poorly correlated to the other metrics, which indicates that they use different features to yield their evaluation (Figure 5).

The annotated resin areas (Figure 3) were also used to evaluate denoised images we obtained using Noise2Void [28] as a way to understand how the processing protocols may perform with machine learning algorithms. After denoising,

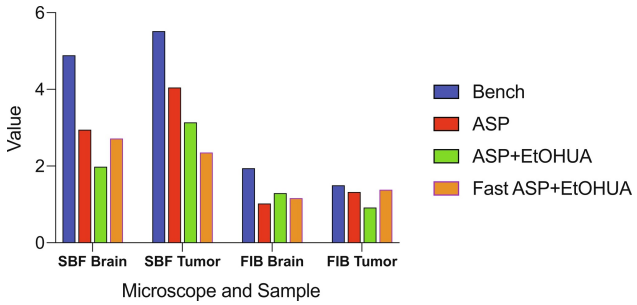


Figure 6: Standard deviation of pixel values in annotated resin areas for brain and tumor samples imaged on the SBF-SEM (SBF) and FIB-SEM (FIB) imaging platforms processed with the bench, ASP-2000 (ASP) ASP-2000 with ethanolic UA (ASP+EtOHUA) or fast ASP-2000 with ethanolic UA (fast ASP+EtOHUA) protocols.

noise values decreased for all samples, regardless of protocol or imaging modality. Noise was reduced by an order of magnitude for FIB-SEM images and by half to 85% for SBF-SEM images while retaining useful signals. Samples imaged with the FIB-SEM had noise values all within one standard deviation regardless of protocol used or tissue type. Surprisingly, after denoising, samples processed with ASP-2000 protocols had less noise than their bench-processed counterparts (Figure 6).

In conclusion, our results indicate that the ASP-2000 automated specimen processor allows for adequate staining of samples and the resulting image quality is suitable for deep learning-based models, such as those used for automated segmentation. This automated processor is capable of being programmed for a variety of samples and protocol needs. Moreover, its utilization decreases the overall time and cost for sample processing, decreases the operator time required, and improves protocol reproducibility.

2.5 Acknowledgments

This research was funded by the OHSU Center for Spatial Systems Biomedicine at Oregon Health and Science University, NIH-NCI Cancer Center Support Grant 2P30CA069533 and OHSU Faculty Initiative Pool Award to CSL. Processing and imaging were performed at the Multiscale Microscopy Core Facility, a University Shared Resource at OHSU. Special thanks to Gordon Mills and Dong Zhang at the OHSU Knight Cancer Institute for providing mouse tissue and Casey Vanderlip and Courtney Glavis-Bloom at Salk Institute for Biological Studies for providing marmoset brain tissue. We also acknowledge support from the M. J. Murdock Charitable Trust for providing funds to purchase the Thermo Scientific Helios 5UC FIB-SEM.

3 Semi-supervised semantic segmentation in Electron Microscopy 3D volumes with sparse labels

3.1 Abstract

Electron microscopy enables imaging at nanometer resolution 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 the 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 chapter, 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, semi-supervised learning as well as next steps for the mitigation of the manual segmentation bottleneck. [2]

3.2 Introduction

Recent advances in cancer nanomedicine have made cancer treatment safer and more effective [34]. Nanotechnology has elucidated interactions between tumor cells and their microenvironment showing key factors in cancer behavior and responses to treatment [35, 36]. 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 [37].

To do this, high resolution information collected from the cellular components at 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 volume Electron Microscopy (vEM) image stacks, and allowing access to 3D information from tissues [38]. 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 [39].

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 annotating. It's a tedious and time consuming task, making it

[2]The work in this chapter was published in Pagano L, Thibault G, Bousselham W, Riesterer JL, Song X, Gray JW. Efficient semi-supervised semantic segmentation of electron microscopy cancer images with sparse annotations. Front Bioinform. 2023 Dec 15;3:1308707

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 [18].

Segmenting images acquired via FIB-SEM is a difficult problem. Indeed, these images differ considerably from natural ones (images representing what human being 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 [40]. Furthermore, the ultrastructure of tumor cells and their microenvironment vary from those of normal cells [41], and EM analysis methods can be tissue type dependent; most current methods have been developed for neural images [42, 43]. 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.

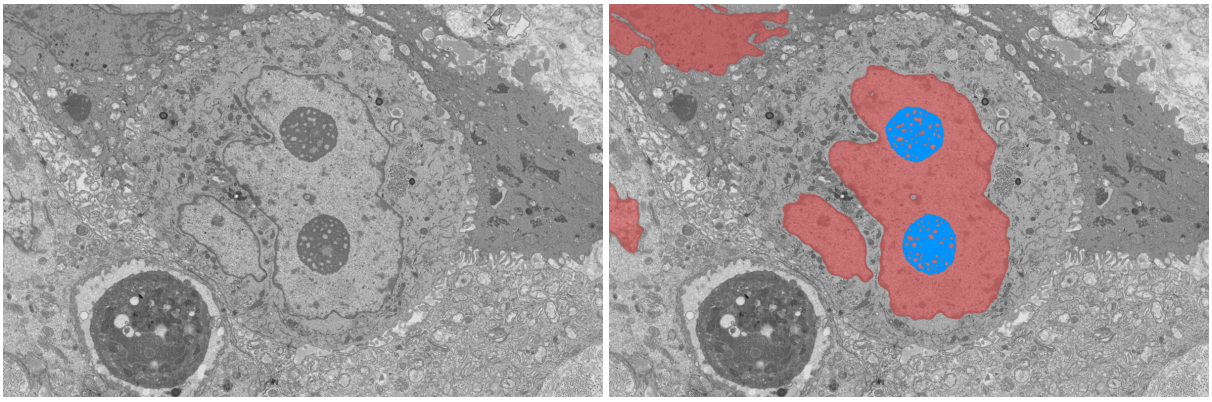
We expanded on previous work from [44], where authors showed that a sparsely manually annotated dataset, typically around 1% of the image stack, was sufficient to train models to segment the whole volume. While state-of-the-art in semantic segmentation has been dominated by attention-based models for natural images [45], convolutional architectures remain main stream with EM data, and were used in [44], 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 chapter, 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 [46] and have emerged as promising therapeutic targets for cancer treatment [47]. 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 7 and 8.

3.3 Materials and methods

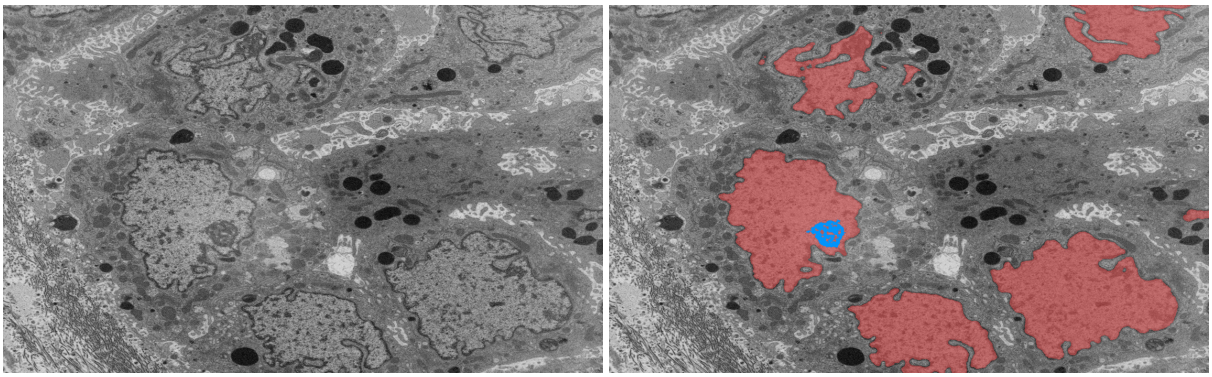
3.3.1 Training and evaluation

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. 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 [44] of extracting tiles of size 2048×2048 and down sampling them to 512×512 as a way to artificially add context. We applied standard random flip and rotation data augmentations. When training with nucleoli,



a Volume 1 example image.

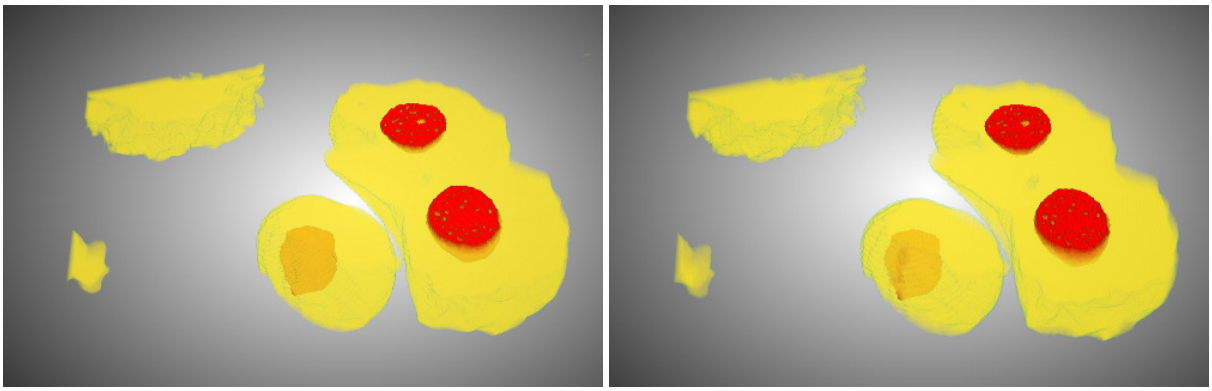
b Volume 1 example image annotations.



c Volume 3 example image.

d Volume 3 example image annotations.

Figure 7: Example image slices and corresponding ground truth annotations from Volumes 1 and 3. The image width is equal to $25\,\mu\mathrm{m}$. 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.



a Volume 1 3D ground truth.

b Volume 1 3D segmentation results.

Figure 8: Ground truth (a) and segmentation by SSL-UNet++-CutMix (b) 3D visualizations for Volume 1. Nuclei are in yellow and nucleoli in red.

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 hours	∼70 hours	∼90 hours	∼144 hours	∼60 hours	∼64 hours

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 classes setting, 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 [44]. Indeed, we found the latter to be biased 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 number of parameters in table 1.

3.3.2 Fully-supervised framework

ResUNet We used previous work from [44] as the baseline for nuclei and nucleoli segmentation. The model used was a Residual U-Net (ResUNet) [48], 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 [49], 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 [50, 51].

UNet++ UNet++ [52] 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 this chapter, 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 [3] 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 $\hat{p}$:

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

FracTALResNet FracTALResNet [53] 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 [4]. It was heavily inspired from 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.

CEECNet CEECNet was also introduced in [53] 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. Motivation behind using this model came from the fact that, as described in section 3.3.1, 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.

SenFormer SenFormer [54] (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.

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 under performed (as shown by the average rank). Detailed results were reported in Tables 3 and 4. For these reasons, it was the model we chose to compare to the baseline in the semi-supervised framework.

[3] `https://github.com/qubvel/segmentation_models.pytorch`
[4] `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

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**

3.3.3 Semi-supervised learning (SSL) framework

Cross Pseudo Supervision (CPS) As described in section 3.3.1, 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 [55]. It trained two networks with a standard supervised cross-entropy loss and used pseudo-labels generated from the segmentation confidence map of one network to supervise the other as can be seen in Figure 9. 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:

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

We trained both ResUNet and UNet++ models with this framework. In this chapter, we trained models with the soft Dice loss defined in equation 1, as we noticed that models trained with a loss closer to the evaluation metric performed

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**

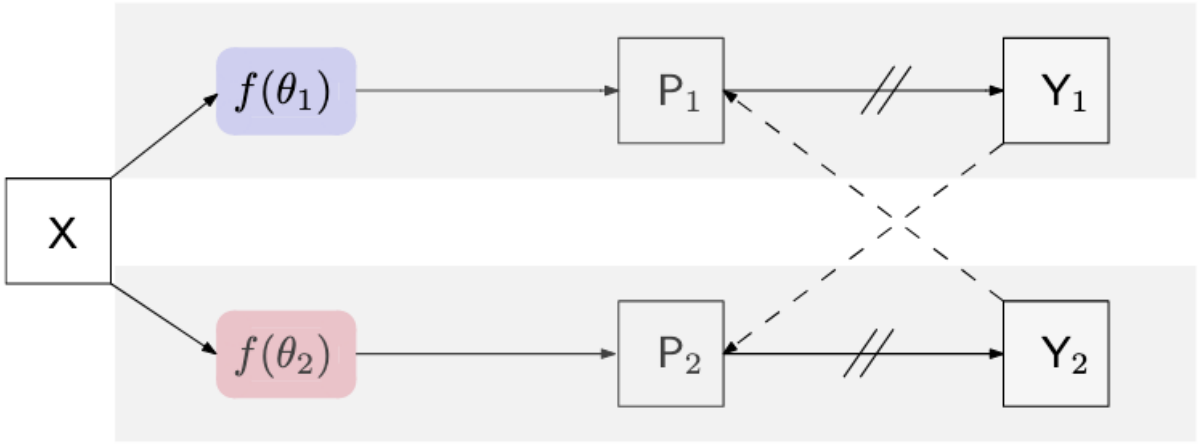
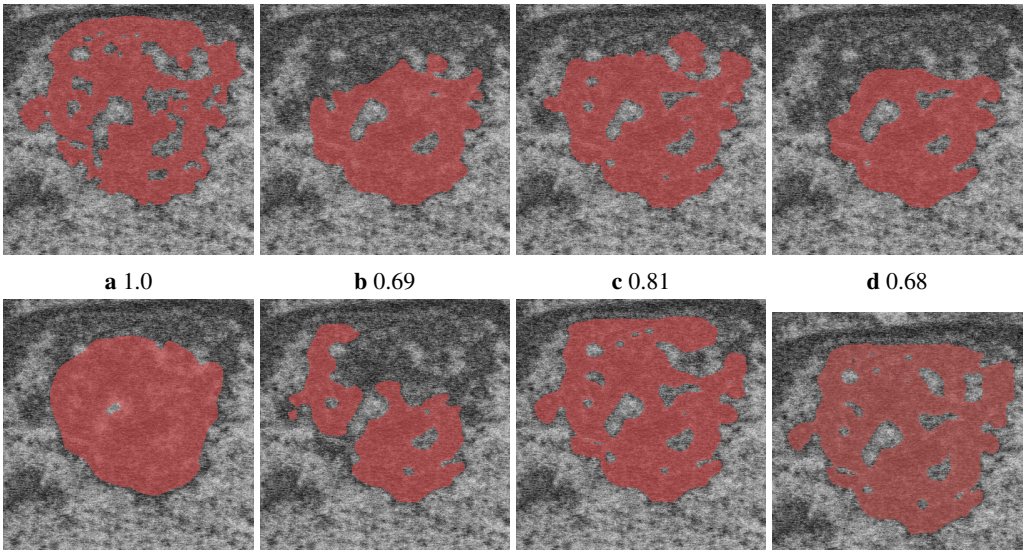


Figure 9: Illustration of the CPS training framework. P_i is the segmentation confidence map, Y_i is the predicted label map. $\rightarrow$ means forward operation, $\dashrightarrow$ loss supervision, and $/\!/$ on $\rightarrow$ stopping the gradient.

better. As a consequence, we replaced all of the cross-entropy losses originally used in [55] by the soft Dice loss. 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 early steps of training. We used a value of 1 for λ in all of our experiments.

Integration of CutMix data augmentation CutMix [56] 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. 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 3.3.1). 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.

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 10. As the



a 1.0 **b** 0.69 **c** 0.81 **d** 0.68

e 0.72 **f** 0.65 **g** 0.89 **h** 0.90

Figure 10: 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. Resolution is $4\,\mathrm{nm}$ per voxel.

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 11), 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 [44] or with supervised models in this chapter.

3.4 Conclusion

In this chapter, 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

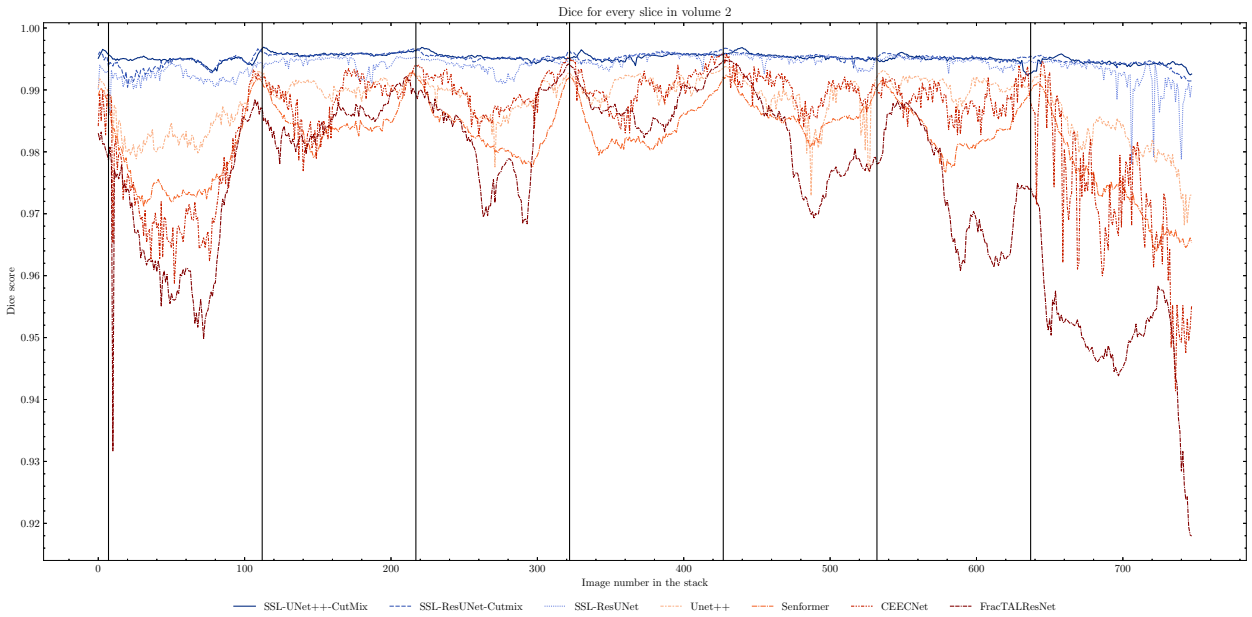


Figure 11: Comparison of Dice scores for nuclei segmentation along all 757 slices in Volume 2 with 7 training images. Training slices are marked with vertical black lines. We can clearly observe the 7 peaks in performance and drops in-between for fully-supervised methods (beige to brown) as opposed to the stability of the SSL models (blue).

Dice on all Volumes. We made the experiment code available at [5] 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.

3.5 Acknowledgments

FIB-SEM data included in this chapter was 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 work is supported by the Cancer Early Detection Advanced Research (CEDAR) Center at the Knight Cancer Institute at OHSU. 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.

3.6 Funding

This chapter 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

[5] `https://github.com/LucasPagano/Segmentation3DEM`

(5U54CA209988), the OHSU Knight Cancer Institute NCI Cancer Center Support Grant (P30CA069533), the OHSU Knight Cancer Institute-Cancer Early Detection and Advanced Research (OHSU) program, and the OCS.

4 Object detection and unsupervised segmentation in large-format, high resolution scanning Electron Microscopy with sparse labels

4.1 Abstract

Electron microscopy is a powerful tool for visualizing cellular structures with high resolution, offering crucial insights into cancer biology. In this chapter, we integrated state-of-the-art object detection and segmentation models with EM to enable accurate identification and quantification of cancer-related organelles. Using Deformable DETR and RetinaNet, we trained models to detect key organelles, including nuclei, nucleoli, mitochondria, and endosomes, in high-resolution scanning electron microscopy images. We applied default data augmentation strategies to improve model robustness and evaluated their performance using mean Average Precison (mAP), Average Recall (AR), and Average Precision at an Intersection over Union (IoU) threshold of 0.5 (AP50). Deformable DETR consistently outperformed RetinaNet across all organelle categories, achieving the highest mAP of 0.683 in nucleus detection and superior recall in all cases. The Segment Anything Model (SAM) was subsequently applied to generate segmentation masks based on the detected bounding boxes, providing access to additional morphology features. While SAM generated high-quality masks, challenges remained in segmenting complex features such as nuclear invaginations and nucleolar fenestrations. Our findings demonstrate the potential of combining AI-driven object detection and segmentation with EM for advancing cancer research. This integration enables precise identification and quantification of sub-cellular structures, facilitating deeper insights into cancer progression, tumorigenesis, and therapeutic responses. The results underscore the importance of using cutting-edge AI models to explore cancer biology at the nanoscale, with implications for the development of personalized cancer treatments.

4.2 Introduction

Cancer, a complex and formidable disease, continues to pose significant challenges to healthcare professionals and researchers worldwide [57]. Despite remarkable strides made in cancer research and treatment, novel approaches that shed light on the intricate mechanisms underlying cancer progression and response to therapy are in constant demand. In recent years, the field of electron microscopy has emerged as a valuable tool for investigating the intricate details of cellular structures at unprecedented resolutions [58, 59]. Among the diverse array of electron microscopy techniques, large format, high resolution scanning electron microscopy stands out as a powerful methodology that offers unique insights into cellular and sub-cellular morphology while scanning enough tissue so that quantification of target organelles can be performed with statistical significance, as opposed to 3D EM which would restrict the covered area too much [60]. This exceptional level of resolution has opened up new avenues for studying cancer-related phenomena at a level of detail that was previously unattainable.

In the context of cancer treatment, EM holds immense promise as a tool for elucidating the underlying mechanisms of tumorigenesis, cancer progression, and therapeutic response [17, 61]. The intricate interplay between cancer cells and their microenvironment and the dynamics of drug interactions within the cellular milieu are all critical factors influencing the efficacy of cancer therapies. EM provides an invaluable platform to visualize and comprehend these intricate processes within cancer cells and their surrounding microenvironment [62].

To further enhance the analysis of EM cancer samples, researchers have turned to the application of object detection artificial intelligence (AI) models [20]. These AI models utilize state-of-the-art deep learning algorithms to identify and delineate structures with remarkable accuracy and efficiency. Traditionally, training these models relied heavily on labeled data, which requires manual annotation and can be time-consuming and costly. However, the advent of semi-supervised learning techniques has opened up new possibilities by leveraging both labeled and unlabeled data. By automating the annotation process, these AI models alleviate the manual burden of identifying and quantifying cellular components within EM images, enabling researchers to analyze large datasets in a more time-effective manner.

The integration of object detection AI models with EM analysis has significant implications for cancer research. By accurately identifying and quantifying nuclei, nucleoli, endosomes, and mitochondria, researchers gain insights into the spatial distribution, density, and morphological alterations of these cellular components in cancer cells [18]. Such information can aid in deciphering the intricate changes associated with tumorigenesis, metastasis, and therapeutic response[62]. Furthermore, the quantitative data generated by these AI models can be leveraged to establish robust correlations between cellular features and clinical outcomes, facilitating the development of personalized cancer treatments.

Moreover, the utilization of AI models in conjunction with EM analysis extends beyond annotation and quantification. These models can assist in pattern recognition, classification, and the identification of rare or abnormal cellular structures within EM images. By automating the identification process, AI models augment the analytical capabilities of researchers and enable the discovery of previously unrecognized cellular phenotypes or sub-cellular alterations associated with cancer [44].

However, the manual annotation of EM images is highly labor-intensive, especially for large datasets where each image can contain thousands of organelles. Traditional deep learning models for organelle segmentation, such as MitoNet [20], rely heavily on dense, pixel-wise annotations, making them less scalable for large-scale studies or for diverse organelle types. To overcome these limitations, we decided to train deep learning models for the detection task, which is way less label-intensive, and rely on unsupervised methods to perform segmentation.

We implemented two state-of-the-art object detection models: RetinaNet [63] and Deformable DETR (DDETR) [64]. RetinaNet was selected for its efficiency in detecting organelles with varying scales, such as endosomes and mitochondria, owing to its focal loss mechanism, balancing detection of small objects against larger ones. On the other hand, DDETR, a transformer-based detection model, was chosen for its capacity to capture complex spatial relationships and

handle irregular organelle shapes, such as nuclei and nucleoli. Its deformable attention mechanism allows the model to learn effectively from limited data, making it ideal for scenarios where labeled samples are sparse.

In conjunction with these detection models, we leveraged the Segment Anything Model (SAM) [21], which enabled us to generate high-quality segmentation masks from bounding boxes produced by our detection models.

The integration of object detection AI models with 2D electron microscopy represents a cutting-edge approach in cancer research. By combining the exceptional resolution of EM with the accuracy and efficiency of AI models, researchers can annotate and analyze cancer samples more effectively. The identification and quantification of nuclei, nucleoli, endosomes, and mitochondria within EM images provide valuable insights into cancer biology, facilitating the understanding of tumorigenesis, metastasis, and therapeutic responses. Managing accurate identification and quantification with few (sparse) labels is a necessary step towards facilitating cancer understanding. This integration holds immense promise for advancing personalized cancer treatments and unlocking new discoveries in the field of cancer research.

4.2.1 Data exploration

The dataset used for this chapter comprises high-resolution $2D$ scanning electron microscopy images (6144×4096 pixels) montaged over large areas , with each montage containing roughly 1000 images. For this analysis, to limit annotation burden, approximately 50 (5%) representative images from each sample were selected for annotations. Meta-data such as cancer subtype, tissue type, nuclear grade and nucleoli fenestration level is also available for each sample.

The images were annotated by our team with CloudFactory, whose annotators underwent specialized training to accurately identify and label organelles of interest, including nuclei, nucleoli, endosomes, and mitochondria. The annotators received detailed guidance to ensure consistency and accuracy in labeling, focusing particularly on the boundaries and morphology of each organelle. Quality Control was performed on both their side and ours. To further refine the annotations, we filtered out bounding boxes smaller than 20 pixels in width or height to remove potential labeling errors or irrelevant detections, ensuring a cleaner dataset for model training.

We have annotated bounding boxes for nuclei, nucleoli, endosomes and mitochondria from 42 samples from different tissue types. This provided us with a total of 131979 mitochondrias, 58489 endosomes, 9875 nuclei and 3727 nucleoli bounding boxes used for training, validation and testing.

4.3 Materials and Methods

4.3.1 Models

Model selection and rationale For this chapter, we selected RetinaNet and Deformable DETR (DDETR) as the primary object detection models. RetinaNet was chosen for its capacity to handle objects of varying scales through

its focal loss mechanism, which is particularly beneficial for detecting small organelles such as endosomes and mitochondria [63]. Deformable DETR was selected for its ability to capture complex spatial arrangements and handle highly irregular organelles like nuclei and nucleoli, as well as its deformable attention mechanism that reduces the need for dense annotations [64].

RetinaNet RetinaNet [63] is a one-stage object detection model based on a Feature Pyramid Network (FPN) architecture. RetinaNet addresses the issue of class imbalance that occurs when background samples vastly outnumber foreground object samples in the training data. To tackle this problem, RetinaNet introduces a novel loss function called the "Focal Loss."

The Focal Loss assigns higher weights to hard examples (i.e., misclassified or challenging objects) during training. By doing so, it emphasizes learning from difficult instances, effectively focusing the model's attention on these problematic samples. This mechanism helps RetinaNet maintain high accuracy even in the presence of heavily imbalanced datasets.

The architecture of RetinaNet comprises a backbone network, which extracts features from the input image, followed by a Feature Pyramid Network. The FPN consists of lateral connections and top-down pathways, which combine features from different levels of the network to capture objects of various sizes. The FPN contributes to the model's ability to detect objects across different scales, making it suitable for multi-object detection tasks.

Deformable DETR Deformable DETR [64] is a state-of-the-art object detection model introduced in 2020. It is an extension of the DETR (DEtection TRansformer) model, which leverages self-attention mechanisms for object detection. The key innovation in Deformable DETR lies in the introduction of deformable attention mechanisms, which enhance the model's spatial modeling capabilities.

In conventional DETR models, attention mechanisms sample fixed regions from feature maps. However, deformable attention allows the model to adaptively sample informative regions, taking into account geometric variations and occlusions in the image. This adaptive sampling significantly improves the model's ability to handle complex scenes and improves the accuracy of object localization.

Deformable DETR further incorporates transformer-based encoders and decoders, allowing it to leverage the benefits of transformer architectures, such as capturing global context and modeling long-range dependencies. The combination of deformable attention and transformer-based modules makes Deformable DETR a highly effective and efficient object detection model.

4.3.2 Large format, high resolution scanning electron microscopy dataset collection

Pre-processing All datasets underwent a standard normalization procedure to ensure consistency in pixel intensity across the entire dataset. Raw electron microscopy images, which inherently vary in intensity, were standardized by adjusting their pixel values to have a mean of 0 and a standard deviation of 1 for each sample. This normalization

was critical for stabilizing the training process and ensuring that the models received uniform inputs, regardless of the original image characteristics.

Following normalization, we performed a filtering step to eliminate erroneous or irrelevant bounding box annotations. Bounding boxes with a width or height of less than 20 pixels were removed from the dataset, as these often resulted from labeling inaccuracies. Manual inspection indicated that such small annotations did not correspond to valid biological structures but were too small to be found during Quality Control.

Given the large size of the original images, roughly 4000×6000 pixels, we applied a slicing procedure to facilitate efficient training and improve detection accuracy, particularly for smaller organelles such as endosomes and mitochondria. Images were divided into smaller patches of 2048×2048 pixels to ensure that the models could process these structures more effectively. This approach was essential for mitigating the memory constraints imposed by GPU training and for optimizing model performance on high-resolution electron microscopy images.

4.3.3 Training and Evaluation

Training Setup Both models were trained on a high-resolution EM dataset that included annotations for four distinct cellular organelles: nuclei, nucleoli, endosomes, and mitochondria. The dataset was split into training, validation, and test sets in an 80/10/10 ratio, respectively. Both models were optimized using the Adam optimizer for the same amount of iterations.

To ensure fairness in model comparisons, all models were trained under the same conditions using four NVIDIA V100 GPUs. Both Deformable DETR and RetinaNet models were trained for the same number of iterations across organelle-specific datasets, with one model trained for each of the four organelles: nuclei, nucleoli, mitochondria, and endosomes. Due to the smaller sizes of endosomes and mitochondria, these models were trained on image slices with dimensions of 2048×2048 pixels, while nuclei and nucleoli models were trained on full size slices.

For data preprocessing, we employed the default data augmentation strategies provided by both Deformable DETR and RetinaNet frameworks. These augmentation techniques included random horizontal and vertical flips, as well as color jittering, to improve the models' robustness and generalization. The input images were normalized using the mean and standard deviation of the training dataset. Each model was initialized with pre-trained weights from the COCO dataset to leverage transfer learning, given the limited size of the annotated EM dataset.

4.3.4 Evaluation Metrics

To evaluate the performance of the models, we used three common metrics in object detection: mAP, Average Recall (AR), and Average Precision at an Intersection over Union (IoU) threshold of 0.5 (AP50). These metrics were computed for each organelle to assess detection accuracy. mAP provides a comprehensive measure of precision across different IoU thresholds, AR evaluates the fraction of objects correctly localized, and AP50 focuses on the precision at a fixed, less harsh IoU threshold of 0.5.

The evaluation was conducted on a holdout test set that was not seen during training, ensuring an unbiased performance assessment. Each model's results were compared across the four organelle categories, enabling a detailed understanding of detection strengths and weaknesses for Deformable DETR and RetinaNet.

4.3.5 Segmentation with Segment Anything Model

Segment Anything Model Segment-Anything Model (SAM) is a deep learning architecture designed for the task of image segmentation [21]. Image segmentation involves dividing an image into different regions or segments to simplify its representation and facilitate analysis. The Segment-Anything Model employs a convolutional neural network (CNN) to process input images and output segmentation masks that delineate different objects or regions within the image.

Unlike traditional image segmentation models that are trained on specific datasets for particular tasks, SAM aims to be versatile and adaptable to various segmentation tasks without the need for extensive retraining.

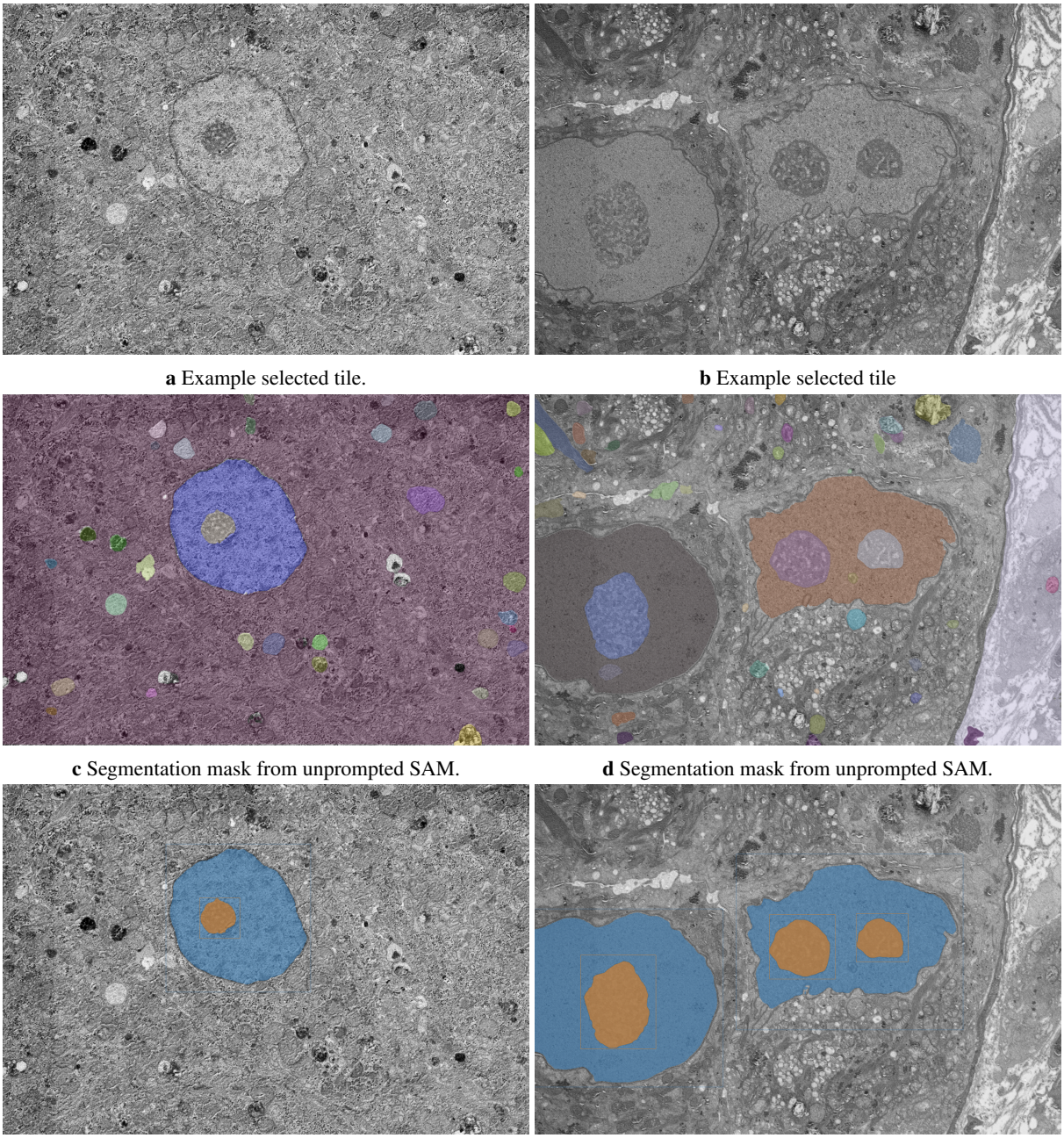
Despite its accrued versatility, as our images differ radically from natural images, even though SAM is able to pick up some structures, the model needs additional information to be able to identify structures present in the images consistently and accurately. As can be seen in Figure 12, unprompted SAM is unable to give meaningful results when fed the entire slice. Another issue is that objects segmented by the unprompted are not identified as belonging to any class, rending downstream analysis of the produced masks very limited.

Bounding box prompting It is possible to prompt SAM with a bounding box region of interest to segment in an image. Since our detector model outputs bounding boxes and SAM accepts these as input to refine its segmentation, we can directly feed the detector bounding boxes to the model to get segmentation masks for detected objects. This considerably improves the accuracy of segmented masks, and comes with the advantage that segmented objects can be classified using the detector class prediction. This prompting technique is critical in producing segmentation masks usable in downstream analysis as can be seen in Figure 12.

4.3.6 Inference with Sliced Aided Hyper Inference

Sliced Aided Hyper Inference (SAHI) [65] is a technique often used to optimize inference in deep learning models, particularly for applications requiring the processing of large, high-resolution images—such as in biological image analysis. The technique helps with managing the computational and memory space challenge faced when analyzing high resolution images by breaking down images into smaller parts, or "slices", before processing. After making predictions on the slices, they are stitched back together to form the full image. The inference results from the slices are reassembled, and overlapping regions can be averaged or combined using additional post-processing steps to ensure a seamless final output.

When studying organelles in Electron Microscopy, SAHI helps the model maintain high accuracy in recognizing small, detailed structures. The key benefits of SAHI in the context of our study are its efficiency, improved accuracy especially for small organelles (endosomes and mitochondria) and scalability.



a Example selected tile.

b Example selected tile

c Segmentation mask from unprompted SAM.

d Segmentation mask from unprompted SAM.

e Segmentation mask from SAM prompted with nuclei (blue) and nucleoli (orange) bounding boxes from detection model.

f Segmentation mask from SAM prompted with nuclei (blue) and nucleoli (orange) bounding boxes from detection model.

Figure 12: Comparison of SAM unprompted and prompted segmentation on example tiles. The image width is equal to $25\,\mu\mathrm{m}$

4.4 Results and Discussion

In this section, we present the results of object detection on Electron Microscopy images for cellular organelles, specifically nuclei, nucleoli, endosomes, and mitochondria. Deformable DETR and RetinaNet were evaluated using mean Average Precison (mAP), Average Recall (AR), and Average Precision at an IoU threshold of 0.5 (AP50). Deformable DETR exhibited superior performance across all organelle categories, with strong results for the detection of nuclei. In particular, we get a mAP of 0.683, which is comparable to results on the natural images COCO detection dataset with full labels. Moreover, our reported AP50 for all organelles outperforms what is reported in state of the art methods such as MitoNet [20] on most of their datasets, especially the cancer images (HeLa). We improved organelle detection AP50 by at least 0.20 mAP which is a considerable improvement.

4.4.1 Results

Nucleus Detection The detection of nuclei, being relatively large and distinct in EM images, yielded the best results across both models. Deformable DETR achieved a significantly higher mAP of 0.683, compared to RetinaNet's 0.474. Additionally, Deformable DETR exhibited a much stronger Average Recall (AR) of 0.901 and AP50 of 0.872, while RetinaNet trailed with an AR of 0.586 and AP50 of 0.728.

The performance gap can be attributed to Deformable DETR's dynamic attention mechanism, which excels at capturing the irregular shapes and large-scale variability of nuclei. The transformer-based architecture effectively focuses on relevant regions, making it particularly suitable for detecting these large, complex organelles, whereas RetinaNet's anchor-based design may struggle with the same level of flexibility and precision in object localization.

Nucleolus Detection Detecting nucleoli proved more challenging. Deformable DETR again demonstrated superior performance, achieving an mAP of 0.524, compared to RetinaNet's 0.298. Deformable DETR also achieved higher scores in terms of AR (0.627) and AP50 (0.793), significantly outperforming RetinaNet's AR of 0.410 and AP50 of 0.524.

The smaller size and irregularities (fenestration, condensation) of nucleoli make their detection difficult, but Deformable DETR's attention mechanism was able to partially overcome these challenges. In contrast, RetinaNet's performance highlights the difficulties faced by anchor-based methods when dealing with small objects in complex environments.

Mitochondria Detection Mitochondria, with their elongated and irregular shapes, posed a challenge for both models, though Deformable DETR again showed better results. It achieved an mAP of 0.494, AR of 0.612, and AP50 of 0.811, while RetinaNet lagged behind with an mAP of 0.300, AR of 0.473, and AP50 of 0.531.

The ability of Deformable DETR to handle object shapes and sizes that vary significantly was evident in these results. Its flexible attention mechanisms allowed for more accurate localization and segmentation of mitochondria, compared to RetinaNet, whose predefined anchors might have been less effective for capturing the elongated structure of mitochondria.

Endosome Detection Endosome detection proved to be the most difficult task for both models, with notably lower performance metrics across the board. Deformable DETR achieved an mAP of 0.365 and AR of 0.543, with an AP50 of 0.771. RetinaNet struggled significantly in comparison, with an mAP of just 0.095, AR of 0.259, and AP50 of 0.266.

The stark difference in performance indicates that Deformable DETR's ability to attend to multiple scales and complex spatial configurations provided a major advantage in detecting these smaller and more ambiguous structures. RetinaNet's lower recall and precision suggest that its anchor-based approach failed to effectively localize and classify endosomes, which are often small and less visually distinctive in EM images.

4.4.2 Discussion

The superior performance of Deformable DETR across all organelle categories is evident from the significantly higher mAP, AR, and AP50 values, especially for the detection of nuclei and mitochondria. Deformable DETR's attention-based mechanism, which can focus dynamically on relevant regions in high-resolution EM images, allowed for better handling of the irregular shapes, varying scales, and complex textures of these organelles. In contrast, RetinaNet's fixed anchor-based detection mechanism struggled to generalize to these variations, particularly for smaller or densely clustered structures like nucleoli and endosomes.

For the larger and more distinct organelles like nuclei, the mAP of 0.683 for Deformable DETR compared to 0.474 for RetinaNet underscores the importance of flexible, scale-aware architectures for detecting objects with significant morphological variation. Similarly, in the detection of mitochondria, which have complex elongated shapes, Deformable DETR's transformer-based architecture outperformed RetinaNet, suggesting its advantage in handling diverse object geometries.

Detection of smaller organelles such as nucleoli and endosomes proved more challenging, particularly for RetinaNet, which achieved notably lower scores. The mAP of 0.095 for endosomes using RetinaNet highlights the limitations of traditional convolutional-based models with fixed receptive fields when dealing with small and less visually distinctive objects.

4.5 Comparative Performance Summary

These results, summarized in table 5 indicate that Deformable DETR consistently outperformed RetinaNet across all organelle categories in terms of mAP, AR, and AP50. The attention-based architecture of Deformable DETR allows for more precise detection of objects with complex and variable morphology, while RetinaNet's anchor-based method is less effective, especially for smaller organelles like endosomes and mitochondria.

4.6 Segmentation Results with Segment Anything Model (SAM)

To further enhance the detection of cellular organelles in EM images, we applied the Segment Anything Model (SAM) [21] to generate segmentation masks from the bounding boxes predicted by Deformable DETR and RetinaNet.

Table 5: Comparative test performance of Deformable DETR and RetinaNet across organelle categories in terms of mAP, AR, and AP50.

Organelle	DDETR mAP / RetinaNet mAP	DDETR AR / RetinaNet AR	DDETR AP50 / RetinaNet AP50
Nucleus	0.683 / 0.474	0.901 / 0.586	0.872 / 0.728
Nucleolus	0.524 / 0.298	0.627 / 0.410	0.793 / 0.524
Mitochondria	0.494 / 0.300	0.612 / 0.473	0.811 / 0.531
Endosome	0.365 / 0.095	0.543 / 0.259	0.771 / 0.266

Due to the absence of ground truth segmentation masks in our dataset, we were unable to perform a quantitative evaluation. However, qualitative assessment of the segmentation results indicated that the masks produced were of high quality overall, effectively capturing the contours and boundaries of cellular structures (see Figure 12). This qualitative success aligns with the broader objective of using AI to automate the analysis of large-scale EM datasets in cancer research.

Despite the generally strong performance, some challenges were noted with specific organelles. SAM encountered difficulties in accurately representing invaginations within the nuclear envelope and fenestration in nucleoli as can be seen in Figure 12f, both of which are morphologically complex structures. These imperfections in segmentation highlight the potential need for further refinement or post-processing steps to accurately capture such nuanced features. Nevertheless, the ability to generate detailed segmentation masks with minimal manual intervention marks a significant step forward in utilizing AI to augment the analysis of EM images, contributing to a deeper understanding of cellular alterations in cancer.

4.7 Conclusion

The integration of advanced artificial intelligence models with electron microscopy represents a transformative approach to studying cellular structures and their alterations in cancer. This research successfully employed Deformable DETR and RetinaNet to detect key organelles (nuclei, nucleoli, mitochondria, and endosomes) in high-resolution EM images. The results demonstrated that Deformable DETR outperformed RetinaNet across all evaluated metrics, particularly excelling in the detection of nuclei, which are critical in understanding tumor biology and therapeutic responses. In particular, we get a mAP of 0.683, which is comparable to results on the natural images COCO detection dataset with full labels. Moreover, our reported AP50 for all organelles outperforms what is reported in state of the art methods such as MitoNet [20] on most of their datasets, especially the cancer images (HeLa) by at least 0.20 mAP, which is a considerable improvement.

Additionally, by leveraging the Segment Anything Model (SAM), we were able to generate high-quality segmentation masks from detection bounding boxes with minimal manual annotations. This approach dramatically reduced the labeling effort required for organelle segmentation, highlighting the potential of combining object detection with prompt-based segmentation methods. While the segmentation results were promising, challenges such as accurately capturing the intricate features of nucleoli and endosomes indicate that there remains room for improvement. Future

refinements could involve optimizing both object detection and segmentation models to better handle the unique morphology of these organelles.

The findings of this chapter underscore the importance of combining high-resolution imaging techniques with sophisticated AI algorithms, paving the way for deeper insights into cancer biology. By accurately identifying and quantifying cellular components, researchers can establish robust correlations between organelle morphology and clinical outcomes, ultimately contributing to the advancement of personalized cancer treatments. As the field continues to evolve, the integration of AI and EM will likely unveil new discoveries that enhance our understanding of cancer and its progression.

4.8 Acknowledgments

FIB-SEM data included in this chapter was 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 work is supported by the Cancer Early Detection Advanced Research (CEDAR) Center at the Knight Cancer Institute at OHSU. 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.

4.9 Funding

This chapter 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), the OHSU Knight Cancer Institute-Cancer Early Detection and Advanced Research (OHSU) program, and the OCS.

5 Summary and Future Work

5.1 Deep learning-based Image Quality Assessment and Denoising

In chapter 2, we introduced a deep learning-based approach to Image Quality Assessment (IQA) and denoising of Electron Microscopy images. We used this approach to evaluate if images produced by protocols with faster process times and less exposure of the operator to hazardous and toxic chemicals and improving the reproducibility of the specimens' heavy metal staining retain the same quality that those of a well-known benchmark protocol. The experimental results showed that faster protocols can indeed be used to collect images of the same quality than those collected with the bench protocol for a variety of samples and protocol needs.

5.2 Semi-supervised semantic segmentation in Electron Microscopy 3D volumes with sparse labels

In chapter 3, we introduced a semi-supervised framework to improve on previous results obtained with the ResUNet architecture for semantic segmentation of nuclei and nucleoli in 3D stacks of Electron Microscopy images of cancer cells. We benchmarked several state of the art fully-supervised models such as UNet++, FracTALResNet, SenFormer, and CEECNet and the semi-supervised Cross Pseudo Supervision framework and drew conclusions on the relative gains of using more complex models, semi-supervised learning as well as next steps for the mitigation of the manual segmentation bottleneck. We gained an insight as to why semi-supervised methods were able to gain as much as 15.6% relative performance increase over fully-supervised models and established guidelines for future work on high variability images such as those of tumor cells.

While the segmentation dice scores obtained in this chapter seem close to perfect, it is important to remember the setting in which we were evaluating our models, which is that a model needs to be trained for each dataset. In the future, we hope that semi-supervised and unsupervised learning techniques can be used to make models able to generalize across samples.

5.3 Object detection and unsupervised segmentation in large-format, high resolution scanning Electron Microscopy with sparse labels

In chapter 4, we introduced a framework for unsupervised segmentation of nuclei, nucleoli, mitochondria and endosomes, using fully-supervised detection. As discussed in section 1.1, while acquiring Electron Microscopy images is now a routine task, their annotation and analysis represents the bottleneck of the image processing pipeline. It is even more true for semantic segmentation annotations, where every pixel in the image needs to be annotated, as opposed to object detection, where an object is only associated with 4 coordinates corresponding to its bounding box. We made use of the ease of annotation of objects in the object detection paradigm to collect a collection of 42 samples with varied

cancer subtypes, tissue types and cancer grades for breast cancer with annotations for the aforementioned organelles. This large volume of annotated data enabled us to train robust detection models and get good detection results with especially strong results for the detection of nuclei. In particular, we report a mAP of 0.683, which is comparable to results on the COCO detection dataset with full labels [19]. Moreover, our reported AP50 for all organelles outperforms what is reported in state of the art methods such as MitoNet [20] on most of their datasets, especially the cancer images (HeLa) by at least 0.20 mAP which is a considerable improvement.

Given the limitations in detecting smaller and less distinctive organelles, further exploration into hybrid architectures or domain-specific enhancements, such as contrast-sensitive loss functions or advanced augmentation techniques, could be investigated to further improve object detection performance in Electron Microscopy images. Additionally, optimizing anchor configurations for RetinaNet or fine-tuning the attention mechanisms in Deformable DETR may help bridge the gap for more challenging object categories. Finding a semi-supervised framework which leverages deformable-DETR would be another way to improve generalizability by making use of the rich information contained in the unlabeled images, which represent 95% of our data, as a large gap still remains between our training and testing performance.

Furthermore, because of encountering instability when training a model on all organelles, our training procedure as of now is to train a detection model for each organelle. However, as the number and granularity of labels increases, it will not be scalable. Thus, a model able to detect all organelles at the same time without decreasing the detection quality is a key next step in establishing a robust large-format, high resolution scanning Electron Microscopy object detection and segmentation pipeline.

5.3.1 Biological constraint

When working with natural images, no assumptions are made about the position of objects to detect or their ability to be superimposed. However, since we are working with subcellular structures, strong assumptions can be made about their relative position, for example, nucleoli can only exist inside nuclei, endosomes and mitochondria can only exist outside nuclei and nucleoli. These assumptions, while trivial, can improve performances thanks to limiting the possible outputs of models at inference and reducing the size of the search space they explore during their training process.

We have experimented with using trained model prediction confidence as a way to sort which bounding boxes to remove first, but we have yet to find a robust setting that improves performance in all cases. We strongly believe that once this post-processing step is established, it will improve models' generalizability and interpretability.

6 References

[1] Gulisa Turashvili and Edi Brogi. Tumor heterogeneity in breast cancer. *Frontiers in Medicine*, 4, December 2017.

[2] Dimitra Georgopoulou et al. Landscapes of cellular phenotypic diversity in breast cancer xenografts and their impact on drug response. *Nature Communications*, 12(1), March 2021.

[3] Lauren A. Hapach, Shawn P. Carey, Samantha C. Schwager, Paul V. Taufalele, Wenjun Wang, Jenna A. Mosier, Nerymar Ortiz-Otero, Tanner J. McArdle, Zachary E. Goldblatt, Marsha C. Lampi, Francois Bordeleau, Jocelyn R. Marshall, Isaac M. Richardson, Jiahe Li, Michael R. King, and Cynthia A. Reinhart-King. Phenotypic heterogeneity and metastasis of breast cancer cells. *Cancer Research*, 81(13):3649–3663, May 2021.

[4] Kevin M. Turner, Syn Kok Yeo, Tammy M. Holm, Elizabeth Shaughnessy, and Jun-Lin Guan. Heterogeneity within molecular subtypes of breast cancer. *American Journal of Physiology-Cell Physiology*, 321(2):C343–C354, August 2021.

[5] Nadia Harbeck, Frédérique Penault-Llorca, Javier Cortes, Michael Gnant, Nehmat Houssami, Philip Poortmans, Kathryn Ruddy, Janice Tsang, and Fatima Cardoso. Breast cancer. *Nature Reviews Disease Primers*, 5(1), September 2019.

[6] Nicholas McGranahan and Charles Swanton. Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. *Cancer Cell*, 27(1):15–26, January 2015.

[7] Juliane M Krüger, Cédric Wemmert, Ludovic Sternberger, Christel Bonnas, Gabriele Dietmann, Pierre Gançarski, and Friedrich Feuerhake. Combat or surveillance? evaluation of the heterogeneous inflammatory breast cancer microenvironment. *The Journal of Pathology*, 229(4):569–578, February 2013.

[8] Q Yang, I Mori, T Sakurai, G Yoshimura, T Suzuma, Y Nakamura, M Nakamura, E Taniguchi, T Tamaki, T Umemura, and K Kakudo. Correlation between nuclear grade and biological prognostic variables in invasive breast cancer. *Breast Cancer*, 8(2):105–110, 2001.

[9] John F. Graf and Maria I. Zavodszky. Characterizing the heterogeneity of tumor tissues from spatially resolved molecular measures. *PLOS ONE*, 12(11):e0188878, November 2017.

[10] Daniele Zink, Andrew Fischer, and Jeffrey Nickerson. Nuclear structure in cancer cells. *Nature reviews. Cancer*, 4:677–87, 10 2004.

[11] Mikael S. Lindström, Deana Jurada, Sladana Bursac, Ines Orsolic, Jiri Bartek, and Sinisa Volarevic. Nucleolus as an emerging hub in maintenance of genome stability and cancer pathogenesis. *Oncogene*, 37(18):2351–2366, 2018.

[12] Matteo Audano, Silvia Pedretti, Simona Ligorio, Maurizio Crestani, Donatella Caruso, Emma De Fabiani, and Nico Mitro. "the loss of golden touch": Mitochondria-organelle interactions, metabolism, and cancer. *Cells*, 9(11):2519, 2020.

[13] Jonathan L. Jeger. Endosomes, lysosomes, and the role of endosomal and lysosomal biogenesis in cancer development. *Molecular Biology Reports*, 47(12):9801–9810, 2020.

[14] Tzipi Cohen Hyams, Keriya Mam, and Murray C Killingsworth. Scanning electron microscopy as a new tool for diagnostic pathology and cell biology. *Micron*, 130(102797):102797, March 2020.

[15] Toshiyuki Ishiwata, Fumio Hasegawa, Masaki Michishita, Norihiko Sasaki, Naoshi Ishikawa, Kaiyo Takubo, Yoko Matsuda, Tomio Arai, and Junko Aida. Electron microscopic analysis of different cell types in human pancreatic cancer spheres. *Oncology letters*, 15(2):2485–2490, 02 2018.

[16] Lorena Signati, Raffaele Allevi, Francesca Piccotti, Sara Albasini, Laura Villani, Marta Sevieri, Arianna Bonizzi, Fabio Corsi, and Serena Mazzucchelli. Ultrastructural analysis of breast cancer patient-derived organoids. *Cancer Cell International*, 21(1):423, 2021.

[17] Jessica L Riesterer, Claudia S López, Erin S Stempinski, Melissa Williams, Kevin Loftis, Kevin Stoltz, Guillaume Thibault, Christian Lanicault, Todd Williams, and Joe W Gray. A workflow for visualizing human cancer biopsies using large-format electron microscopy. *Methods Cell Biol.*, 158:163–181, March 2020.

[18] Alex J. Perez, Mojtaba Seyedhosseini, Thomas J. Deerinck, Eric A. Bushong, Satchidananda Panda, Tolga Tasdizen, and Mark H. Ellisman. A workflow for the automatic segmentation of organelles in electron microscopy image stacks. *Frontiers in Neuroanatomy*, 8, 11 2014.

[19] Papers with code. Papers with code object detection on coco dataset, 2024.

[20] Ryan Conrad and Kedar Narayan. Instance segmentation of mitochondria in electron microscopy images with a generalist deep learning model trained on a diverse dataset. *Cell Systems*, 14(1):58–71.e5, 2023.

[21] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, et al. Segment anything. *arXiv:2304.02643*, 2023.

[22] Yunfeng Hua, Philip Laserstein, and Moritz Helmstaedter. Large-volume en-bloc staining for electron microscopy-based connectomics. *Nature Communications*, 6(1):7923, 2015.

[23] Alain Horé and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th International Conference on Pattern Recognition*, pages 2366–2369, 2010.

[24] Melainia McClain, Stephanie H Nowotarski, Xia Zhao, and Alejandro Sánchez Alvarado. Rapid automated en bloc staining for sem of sections. In *Microscopy and Microanalysis Conference*, 2017.

[25] Shawn Mikula, Jonas Binding, and Winfried Denk. Staining and embedding the whole mouse brain for electron microscopy. *Nature Methods*, 9(12):1198–1201, 2012.

[26] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21:4695–4708, 2012.

[27] Li Sze Chow and Raveendran Paramesran. Review of medical image quality assessment. *Biomedical Signal Processing and Control*, 27:145–154, 5 2016.

[28] Alexander Krull, Tim-Oliver Buchholz, and Florian Jug. Noise2void - learning denoising from single noisy images. *CoRR*, abs/1811.10980, 2018.

[29] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[30] Z. Wang, E.P. Simoncelli, and A.C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402 Vol.2, 2003.

[31] Lin Zhang and Hongyu Li. Sr-sim: A fast and high performance iqa index based on spectral residual. In *2012 19th IEEE international conference on image processing*, pages 1473–1476. IEEE, 2012.

[32] Wufeng Xue, Lei Zhang, Xuanqin Mou, and Alan C. Bovik. Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *IEEE Transactions on Image Processing*, 23(2):684–695, 2014.

[33] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *CoRR*, abs/2004.07728, 2020.

[34] J. Shi, P. W. Kantoff, R. Wooster, and O. C. Farokhzad. Cancer nanomedicine: progress, challenges and opportunities. *Nat Rev Cancer*, 17(1):20–37, 01 2017.

[35] E. Hirata and E. Sahai. Tumor Microenvironment and Differential Responses to Therapy. *Cold Spring Harb Perspect Med*, 7(7), Jul 2017.

[36] H. Y. Tanaka and M. R. Kano. Stromal barriers to nanomedicine penetration in the pancreatic tumor microenvironment. *Cancer Sci*, 109(7):2085–2092, Jul 2018.

[37] R. Baghban, L. Roshangar, R. Jahanban-Esfahlan, K. Seidi, A. Ebrahimi-Kalan, M. Jaymand, S. Kolahian, T. Javaheri, and P. Zare. Tumor microenvironment complexity and therapeutic implications at a glance. *Cell Commun Signal*, 18(1):59, 04 2020.

[38] Lucille A. Giannuzzi and Fred A Stevie. Introduction to focused ion beams. 2005.

[39] Andrew J Bushby, Kenneth M Y P, Robert D Young, Christian Pinali, Carlo Knupp, and Andrew J Quantock. Imaging three-dimensional tissue architectures by focused ion beam scanning electron microscopy. 2011.

[40] Cefa Karabağ, Martin L. Jones, Christopher J. Peddie, Anne E. Weston, Lucy M. Collinson, and Constantino Carlos Reyes-Aldasoro. Semantic segmentation of hela cells: An objective comparison between one traditional algorithm and four deep-learning architectures. *PLoS ONE*, 15, 10 2020.

[41] Leonard Nunney, Carlo C. Maley, Matthew Breen, Michael E. Hochberg, and Joshua D. Schiffman. Peto's paradox and the promise of comparative oncology. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370, 7 2015.

[42] Abdel Aziz Taha and Allan Hanbury. Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. *BMC Medical Imaging*, 15:29, 2015.

[43] Fan Zhang, Anna Breger, Kang Ik Kevin Cho, Lipeng Ning, Carl Fredrik Westin, Lauren J. O'Donnell, and Ofer Pasternak. Deep learning based segmentation of brain tissue from diffusion mri. *NeuroImage*, 233:117934, 6 2021.

[44] Archana Machireddy, Guillaume Thibault, Kevin G. Loftis, Kevin Stoltz, Cecilia E. Bueno, Hannah R. Smith, Jessica L. Riesterer, Joe W. Gray, and Xubo Song. Segmentation of cellular ultrastructures on sparsely labeled 3d electron microscopy images using deep learning. *Frontiers in Bioinformatics*, 3, 2023.

[45] Jitesh Jain, Anukriti Singh, Nikita Orlov, Zilong Huang, Jiachen Li, Steven Walton, and Humphrey Shi. Semask: Semantically masked transformers for semantic segmentation. 12 2021.

[46] Daniele Zink, Andrew H. Fischer, and Jeffrey A. Nickerson. Nuclear structure in cancer cells. *Nature reviews. Cancer*, 4:677–687, 9 2004.

[47] Mikael S. Lindström, Deana Jurada, Sladana Bursac, Ines Orsolic, Jiri Bartek, and Sinisa Volarevic. Nucleolus as an emerging hub in maintenance of genome stability and cancer pathogenesis. *Oncogene*, 37(18):2351–2366, 2018.

[48] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[49] Xavier Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. *Journal of Machine Learning Research - Proceedings Track*, 9:249–256, 01 2010.

[50] Run Su, Deyun Zhang, Jinhuai Liu, and Chuandong Cheng. Msu-net: Multi-scale u-net for 2d medical image segmentation. *Frontiers in Genetics*, 12:140, 2 2021.

[51] Nahian Siddique, Sidike Paheding, Colin P. Elkin, and Vijay Devabhaktuni. U-net and its variants for medical image segmentation: theory and applications. *IEEE Access*, 11 2020.

[52] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. *CoRR*, abs/1807.10165, 2018.

[53] Foivos I. Diakogiannis, François Waldner, and Peter Caccetta. Looking for change? roll the dice and demand attention. *Remote Sensing 2021, Vol. 13, Page 3707*, 13:3707, 9 2021.

[54] Walid Bousselham, Guillaume Thibault, Lucas Pagano, Archana Machireddy, Joe Gray, Young Hwan Chang, and Xubo Song. Efficient self-ensemble for semantic segmentation, 2022.

[55] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision, 2021.

[56] Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Seong Joon Oh, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. *Proceedings of the IEEE International Conference on Computer Vision*, 2019-October:6022–6031, 5 2019.

[57] Rebecca L Siegel, Angela N Giaquinto, and Ahmedin Jemal. Cancer statistics, 2024. *CA Cancer J Clin*, 74(1):12–49, Jan-Feb 2024.

[58] Naava Naslavsky and Steve Caplan. The enigmatic endosome - sorting the ins and outs of endocytic trafficking. *J. Cell Sci.*, 131(13), July 2018.

[59] Makoto Abe and Nobuhiko Ohno. Recent advancement and human tissue applications of volume electron microscopy. *Microscopy (Oxf)*, Oct 2024.

[60] Anusha Aswath, Ahmad Alsahaf, Ben N.G. Giepmans, and George Azzopardi. Segmentation in large-scale cellular electron microscopy with deep learning: A literature survey. *Medical Image Analysis*, 89:102920, 2023.

[61] N G Ordóñez and B Mackay. Electron microscopy in tumor diagnosis: indications for its use in the immunohistochemical era. *Hum Pathol*, 29(12):1403–1411, Dec 1998.

[62] Jessica L Riesterer, Cecilia Bueno, Erin S Stempinski, Steven K Adamou, Claudia S López, Guillaume Thibault, Lucas Pagano, Joseph Grieco, Samuel Olson, Archana Machireddy, et al. Large-Scale Electron Microscopy to Find Nanoscale Detail in Cancer. *Microscopy and Microanalysis*, 29(Supplement_1):1078–1079, 07 2023.

[63] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007, 2017.

[64] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

[65] Fatih Cagatay Akyon, Sinan Onur Altinuc, and Alptekin Temizel. Slicing aided hyper inference and fine-tuning for small object detection. In *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, October 2022.