From Natural to Nanoscale: Training ControlNet on Scarce FIB-SEM Data for Augmenting Semantic Segmentation Data

Hannah Kniesel*
Visual Computing Group
Ulm University
hannah.kniesel@uni-ulm.de

Pascal Rapp*
Visual Computing Group
Ulm University
pascal.rapp@uni-ulm.de

Pedro Hermosilla
Computer Vision Lab
TU Vienna
phermosilla@cvl.tuwien.ac.at

Timo Ropinski
Visual Computing Group
Ulm University
timo.ropinski@uni-ulm.de

Abstract

Focused Ion Beam Scanning Electron Microscopy (FIB-SEM) is widely used for ultrastructural imaging, with segmentation of FIB-SEM stacks being essential for downstream quantification and biological analysis. However, manual annotation of these datasets is labor-intensive and time-consuming. Training semantic segmentation models offers a scalable alternative, but FIB-SEM datasets are typically small and exhibit low variance as the imaging process involves slicing and capturing individual sample sections. This poses a significant challenge for model training.

Data augmentation via generative models has emerged to address limited data, with diffusion models showing state-of-the-art synthesis capabilities. Yet, their dependence on large natural image datasets restricts direct application to FIB-SEM data. In this work, we explore fine-tuning ControlNet – a conditional diffusion model extension – on small FIB-SEM datasets to produce realistic, label-consistent synthetic images for segmentation. Despite relying on a diffusion backbone trained exclusively on natural images, we show that fine-tuning ControlNet with domain-specific structural cues enables effective data augmentation, leading to an impressive downstream mIoU improvement of up to $+15.4$. We compare ControlNet augmentations against standard augmentation techniques in respect to generation time as well as downstream task performance. We additionally explore different dataset sizes, and provide insights into the feasibility of applying large-scale generative models in data-scarce, low-variance scientific imaging domains like FIB-SEM.

*These authors contributed equally to this work.

1. Introduction

EM has become an indispensable tool in materials science and biology for studying structures at the nanoscale [2, 14, 29, 31]. Sepcifically, Focused Ion Beam Scanning Electron Microscopy (FIB-SEM), a specialized sub-modality of EM, is increasingly vital for ultrastructural imaging of complex biological architectures, such as neuronal circuits in connectomics or cellular organelles [8, 22, 23]. Crucially, this technique generates data by iteratively milling away nanometer-thick layers of a sample and imaging the freshly exposed surface. The comprehensive data analysis and quantification of these complex FIB-SEM datasets, critically relies on the accurate segmentation of the imaged structures [28, 38]. However, performing this segmentation manually is exceedingly time-consuming and often subjective, making it a major bottleneck for large-scale studies [30, 35]. Hence, there is a strong imperative to train automated deep learning models for this task. Yet, collecting large-scale datasets for machine learning tasks such as image segmentation or classification remains a significant challenge. The process of acquiring EM images is both time-consuming and resource-intensive [16, 37], requiring access to specialized instrumentation and sample preparation protocols that must be carried out by trained experts. Furthermore, image annotations – especially for semantic segmentation – typically demand domain expertise, making large annotated datasets prohibitively expensive and slow to assemble. Unfortunately, neural networks such as CNNs often struggle when trained on limited data [41]. In FIB-SEM datasets, the low variance – due to high similarity between adjacent slices – further compounds this issue. The limited effective diversity restricts the model's exposure to distinct visual features, making it particularly challenging to train

robust segmentation models without overfitting.

To address the limited availability of training data in EM, data augmentation has emerged as a key research area. Traditional augmentation techniques, such as random cropping, flipping, or contrast adjustments, have been widely used [37, 39, 39, 44], but are inherently limited in the diversity they introduce. Hence, generative models, such as Generative Adversarial Networks (GANs) [12], have shown promise in generating synthetic EM-like images for use in data augmentation pipelines [36]. These models offer a cost-effective way to enrich training datasets compared to the substantial expense of additional image acquisition via EM instruments.

With the advent of diffusion models, generative image synthesis has reached new state-of-the-art levels in terms of realism, diversity, and control [9, 17, 18, 32, 40, 48]. These models, particularly Stable Diffusion [32], have demonstrated remarkable capabilities in generating high-quality natural images with feasible compute resources, due to computations in a low dimensional latent space. Hence, these models have rapidly become the foundation for a new generation of data-driven applications [25, 45, 46]. However, adapting these models to specialized imaging domains like EM presents substantial challenges. Diffusion models are generally trained on massive datasets of natural images, and their performance degrades considerably when applied to domains with very different statistical properties, such as grayscale, high-frequency, and noise-prone EM imagery.

This domain shift renders pretrained diffusion models largely ineffective when used out of the box for EM data generation. Additionally, although training diffusion models does not necessarily require labeled data, it still demands a large quantity of raw images, which remains a limiting factor in the EM domain due to the high acquisition cost. Consequently, diffusion-based data augmentation for EM has yet received limited attention, despite its promising potential.

In this paper, we investigate whether diffusion models pretrained on natural images can be leveraged for FIB-SEM data augmentation using only a small number of domain-specific samples. Specifically, we explore fine-tuning ControlNet [48] – a conditioning mechanism for diffusion models – on small FIB-SEM datasets to guide image synthesis in a controllable way. This approach is particularly well-suited for segmentation tasks, as ControlNet enables conditioning on structural cues such as segmentation masks, allowing the generation of image–label pairs. This, in turn, allows us to augment limited datasets with synthetic data to improve downstream model performance. Our contributions are as follows:

- We show that ControlNet, when fine-tuned on a small FIB-SEM dataset, can generate structurally coherent images conditioned on segmentation masks, even when using a Stable Diffusion backbone pretrained exclusively on natural images.
- We evaluate the generated data in the context of semantic segmentation, demonstrating measurable improvements when augmenting training sets with ControlNet-generated samples.
- We compare our approach to standard data augmentation techniques and assess the trade-offs in generation time and sample utility under varying dataset sizes.

By doing so, we aim to make a first step toward bringing state-of-the-art generative modeling into data-scarce domains like EM.

2. Related Work

This section provides an overview of existing literature relevant to our work, structured around generative models in scientific imaging, the advent of conditional diffusion models, and data augmentation techniques focusing on EM and other data-scarce scientific domains.

Semantic Segmentation for Biological EM Semantic segmentation is fundamental to quantitative analysis in biological EM. It underpins critical scientific discoveries, enabling detailed morphological studies of cellular structures, organelles, membranes, and neuronal circuits [2, 29, 31]. Historically, EM image segmentation relied on laborious manual annotation or computationally intensive classical image processing techniques such as thresholding, watershed algorithms [3], or active contours [20]. While providing some level of automation, these methods often struggled with the complex, noisy, and highly varied morphology inherent to EM data, requiring extensive manual correction.

The advent of deep learning, particularly CNNs, revolutionized medical and biological image analysis, including EM segmentation. Architectures like the U-Net [33], with its symmetric encoder-decoder structure and skip connections, quickly became the defacto standard due to their exceptional performance on biomedical imaging tasks, even with relatively small datasets. Subsequent adaptations and enhancements of U-Net [5, 26, 34], such as 3D U-Nets for volumetric EM data [24] and densly-connected variants [6], have further improved performance in various EM contexts. These deep learning models are capable of learning intricate, high-level features directly from raw image data, leading to significantly more accurate and efficient segmentation compared to traditional methods.

Generative Models The scarcity of large, annotated datasets in specialized scientific domains, including EM, has motivated the application of generative models for data augmentation. Early efforts often leveraged GANs [12] to synthesize domain-specific images [15, 19, 21, 36, 42].

GANs have for instance been used to generate synthetic EM-like images for augmenting training pipelines, showing promise in tasks like Herpesvirus detection [36]. These models offered a cost-effective alternative to expensive real image acquisition. However, GANs often suffer from training instability, mode collapse, and difficulty in controlling the semantic content of generated images.

More recently, diffusion models have emerged as state-of-the-art generative models, demonstrating unprecedented realism, diversity, and controllable synthesis capabilities across various natural image tasks [9, 17, 18, 32, 40]. Their success has naturally led to their exploration in scientific and medical imaging. For example, Wu et al. [45] introduced MedSegDiff, a diffusion model specifically designed for medical image segmentation. Similarly, works like SegGen [46] and studies on open-vocabulary object segmentation [25] demonstrate the power of diffusion models in generating images with corresponding masks or understanding novel semantic concepts, respectively. In the EM domain itself, dedicated diffusion models such as EMDiffuse have been proposed for tasks like denoising and super-resolution by fine-tuning on noisy-clean image pairs [27]. While these applications show the versatility of diffusion models in scientific contexts, many still require substantial domain-specific data or are tailored to specific tasks other than semantic segmentation augmentation with explicit label control.

Conditioning in Diffusion Models The ability to exert fine-grained control over the image generation process is a critical feature for practical applications of generative models, allowing to not only generate synthetic data, but to also create labeled synthetic data. Conditional diffusion models achieve this by integrating external inputs that guide the synthesis. A pivotal advancement in this area is ControlNet [48], which revolutionized the application of large, pre-trained text-to-image diffusion models like Stable Diffusion [32]. ControlNet works by cloning the encoder of a pre-trained diffusion model and training it to accept various forms of conditioning signals – such as edge maps, segmentation masks, depth maps, or keypoints – while keeping the original diffusion model's weights frozen. This architecture allows ControlNet to leverage the vast, high-quality generative knowledge embedded in the frozen backbone (trained on natural images) and adapt it to conditional inputs with relatively small datasets. In this work, we investigate not only the ability of ControlNet to follow conditional inputs after fine-tuning on a small EM segmentation dataset, but also whether the frozen Stable Diffusion backbone, originally trained on natural images, can effectively handle the domain shift introduced by EM data.

Beyond its initial applications in natural image editing and controlled generation, ControlNet has begun to find traction in diverse biomedical imaging applications. Examples include adaptive whole-body PET image denoising [47] and applications in X-ray image synthesis [11]. The success of these applications underscores ControlNet's versatility in adapting powerful generative models to specialized image modalities and specific tasks. Our work extends this promising direction by investigating ControlNet's capacity to generate label-consistent EM images for semantic segmentation, directly addressing the challenge of data scarcity in this unique scientific domain by leveraging structural cues from semantic masks.

Data Augmentation for Small Datasets The challenge of limited data is pervasive in scientific machine learning, particularly in high-cost imaging modalities like EM [16, 37]. Neural networks, such as CNNs commonly used for segmentation, are known to perform poorly when trained on insufficient data [41]. To mitigate this, data augmentation is a crucial technique. Traditional augmentation methods, including geometric transformations (e.g., random cropping, flipping, rotation) and intensity adjustments (e.g., contrast, brightness) are widely applied in EM and other fields [39, 44]. While effective, these methods are inherently limited in the novel diversity they introduce, as they only apply transformations to existing samples rather than generating genuinely new content.

Generative models, as discussed above, represent a more advanced form of data augmentation capable of synthesizing entirely new samples. Earlier works in EM have explored GANs for this purpose, demonstrating their potential to enrich training datasets [36]. However, these generative approaches for EM data augmentation not only faced limitations regarding image quality, diversity, but more importantly, the precise semantic control required for tasks like semantic segmentation. Our work directly addresses these limitations by introducing ControlNet-based augmentation for EM. By leveraging ControlNet's ability to generate images conditioned on explicit segmentation masks, we can produce synthetic EM image-label pairs that are consistent with the desired segmentation, offering a superior and more controlled form of data enrichment compared to prior methods. This represents a significant step towards improving downstream task performance of semantic segmentation even in data-scarce scientific contexts.

3. Data Challenges in EM

EM data for semantic segmentation differs significantly from natural image datasets like ADE20k [49], Pascal VOC [10], or COCO-Stuff [4], which presents unique challenges and opportunities for generative data augmentation.

Firstly, EM images are fundamentally distinct in appearance from RGB natural images that typically depict object surfaces. EM encompasses various modalities,

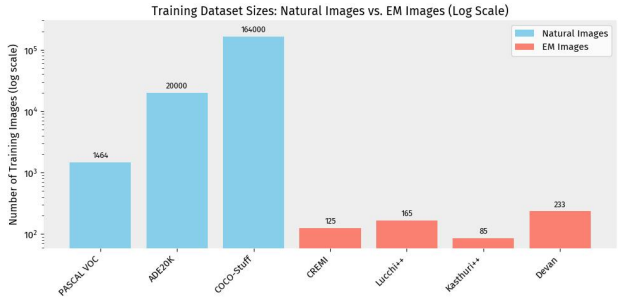

Figure 1. Training dataset sizes for common semantic segmentation benchmarks, comparing natural image datasets with those used in electron microscopy (EM). Natural image datasets include PASCAL VOC [10], ADE20K [49], and COCO-Stuff [4]. EM datasets include CREMI [1], Lucchi++ and Kasthuri++[7], as well as Dataset 1 from Devan et al.[37]. Note that the y-axis is shown on a logarithmic scale for improved visibility across the wide range of dataset sizes. The chart highlights the significantly smaller scale of available EM training data.

each offering unique perspectives: Scanning Electron Microscopy (SEM) visualizes sample surfaces; Transmission Electron Microscopy (TEM) provides 2D projections of internal structures by transmitting electrons through ultrathin sections; Scanning Transmission Electron Microscopy (STEM) offers high-resolution insights by scanning a focused beam through the sample; and FIB-SEM combines ion milling with SEM imaging to acquire serial sections. In this work, we specifically focus on FIB-SEM data, which, like most EM modalities, is grayscale and often reveals the intricate internal ultrastructure of samples at nanometer resolutions. The visual content in these images is highly specialized, frequently repetitive (e.g., membrane bilayers), and follows specific biological rules, diverging significantly from the diverse semantics of natural scenes. Additionally, EM images often contain unique forms of noise and artifacts due to the specialized acquisition and preparation processes. This profound domain gap necessitates fine-tuning generative models to synthesize realistic EM-specific imagery.

Secondly, EM datasets are typically much smaller (see Figure 1). This is due to the high cost, time-consuming nature, and specialized expertise required for both image acquisition (e.g., expert instrument setup) and precise annotation. This data scarcity directly challenges the applicability of large-scale diffusion models, which conventionally rely on vast training corpora.

Thirdly, FIB-SEM datasets for semantic segmentation mostly originate from image stacks [1, 7, 13, 37], and hence exhibit limited variance. Individual images within a stack often show minimal differences, leading to high correlation. This low variance can hinder standard augmentation techniques, which might inadvertently distort the underlying data distribution, potentially misdirecting model capacity. Conversely, training a generative model offers a unique opportunity to learn directly from this underlying distribution, enabling the synthesis of new, varied samples that remain faithful to the domain's inherent statistical properties. Within our work we focus on these common low-variance datasets and leverage FIB-SEM data from [37] (details see subsection 4.1).

4. Method

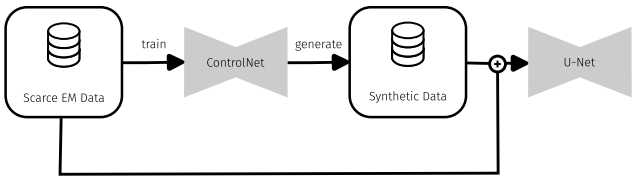

Figure 2. The method of our work: We finetune ControlNet on a small EM dataset with segmentation masks. We then use the trained ControlNet to generate a set of synthetic images, which we use for enhancing the real dataset for training a downstream segmentation model (U-Net).

Our approach focuses on adapting a pre-trained ControlNet model with a frozen Stable Diffusion backbone, pretrained on natural images, for the synthesis of realistic, label-consistent FIB-SEM images, specifically targeting data augmentation for semantic segmentation tasks with scarce data (see Figure 2). This section details the dataset, training strategy, image generation and downstream model training process employed in our work.

4.1. Dataset

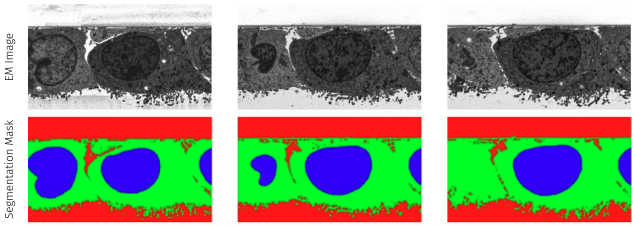

Figure 3. Representative samples from the dataset along with their corresponding color-coded segmentation masks used to condition ControlNet. The visual consistency across samples highlights the low variability within the dataset.

We utilize a FIB-SEM dataset [37] for training and evaluating both ControlNet and downstream segmentation models. To generate this dataset, embedded cells were mounted in a scanning electron microscope (SEM) equipped with a focused ion beam (FIB). The FIB sequentially removes thin layers from the sample, and after each ablation, the newly

exposed surface is imaged using the SEM. Repeating this process hundreds of times yields a series of high-resolution images that can be reconstructed into a three-dimensional (3D) volume [43].

Each image has a resolution of 896×512 pixels and is annotated with three semantic classes: Background, Cytoplasm, and Nucleus (see Figure 3). The dataset of real EM images was partitioned into training, validation, and test sets, comprising 233, 25, and 65 images, respectively. Notably, the validation and test sets contain only real EM images to ensure an unbiased evaluation of model generalization. All data augmentation and enhancement methods were applied solely to the training set.

4.2. ControlNet

We leverage the ControlNet framework as introduced by Zhang et al. [48]. For our experiments, we utilize the pre-trained Stable Diffusion [32] 1.5 backbone, which has demonstrated remarkable capabilities in generating high-quality natural images.

A primary challenge in applying a natural image-trained backbone to EM data is the domain shift represented in the image dimensions: Stable Diffusion models typically operate on 3-channel RGB images, whereas EM images are inherently grayscale (single channel). To address this, and to limit the amount of changes made to the model itself, we adapted our single-channel EM images by replicating the grayscale channel three times to create a pseudo-RGB image, thereby matching the expected dimension of the Stable Diffusion 1.5 backbone. This straightforward approach allows the pre-trained model to process EM images without architectural modifications to its core components.

4.2.1. Conditional Control with Semantic Segmentation Masks

To enable the generation of label-consistent synthetic images for semantic segmentation, we follow the conditional training strategy for ControlNet using color-encoded semantic segmentation masks, as described in the original work [48]. Given our dataset's three distinct biological classes (for details see subsection 4.1), we encode each class within a specific channel of a 3-channel RGB mask: the first class is encoded in the Red channel, the second in the Green, and the third in the Blue channel. While this allows for three primary class representations, it's crucial to understand that the RGB format itself can encode a multitude of colors by blending these channels, thereby enabling the representation of more than three distinct classes, hence allowing the method to be extended to a multitude of classes. This explicit encoding provides direct, structural cues to the ControlNet model, guiding the generative process to produce images that precisely correspond to the provided segmentation map. The ability to condition generation on semantic information is paramount for creating valuable synthetic image-label pairs for downstream segmentation model training.

4.2.2. Textual Prompts for Domain Guidance

Beyond structural conditioning, textual prompts play a vital role in steering the generative process, especially when adapting models from natural to specialized domains. For each training data point, a random prompt is selected from a predefined list of 37 unique descriptions pertaining to various cellular components and structures observed in EM. These text prompts were meticulously formulated based on representative images within our EM dataset. An iterative process of adjustment was undertaken by testing these prompts with the non-fine-tuned Stable Diffusion backbone, ensuring that they provided a reasonable starting point for further fine-tuning. Care was taken to formulate prompts that describe structures and color values as closely as possible to the target EM data, even prior to fine-tuning (examples see Figure 4 Epoch 0).

During this preliminary phase, we identified a significant domain gap: the general-purpose Stable Diffusion backbone, even when guided by textual prompts, frequently generated images resembling conventional SEM. These SEM-like outputs primarily depicted surface-level structures of the samples – visual characteristics that align more closely with natural images. In contrast, our dataset consists of FIB-SEM images, which visualize internal cellular structures markedly different from natural image distributions.

This discrepancy is qualitatively illustrated in Figure 4 (Epoch 0), where the generated output exhibits typical SEM-style surface textures rather than the expected cross-sectional views seen in FIB-SEM. These observations highlighted the need to fine-tune ControlNet specifically for this domain to bridge the pronounced visual mismatch and enable faithful reproduction of the internal structures characteristic of FIB-SEM imaging.

4.2.3. Synthetic Data Generation

After fine-tuning ControlNet on our small dataset, we generate a set of synthetic EM images for data augmentation. This generation process mirrors the conditional setup used during training: real segmentation masks from our dataset, encoded as RGB images, serve as the primary structural conditioning input. For each generated image, a text prompt is randomly sampled from the same pool of 37 prompts used during the fine-tuning phase. This ensures that the generated synthetic data benefits from both the explicit structural guidance of the segmentation mask and the general textural and semantic cues embedded in the learned prompt representations, facilitating the creation of diverse yet label-consistent EM imagery.

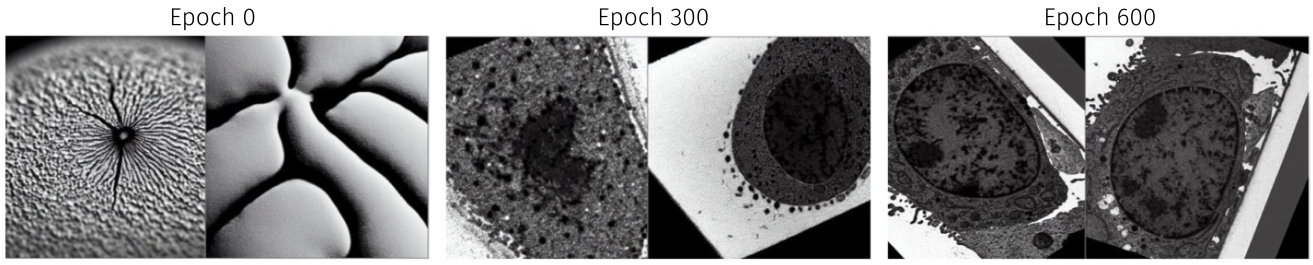

Figure 4. Training progression of ControlNet. Initially, the model is unable to generate realistic FIB-SEM images, but it gradually converges toward increasingly accurate and realistic representations as training progresses.

4.3. U-Net Training for Semantic Segmentation

To evaluate the efficiency of our ControlNet-generated synthetic EM images, we proceed with training a standard U-Net model [33] for semantic segmentation. We construct the final training dataset by combining varying proportions of real EM images with the newly generated synthetic images. During training, we explore various scenarios by mixing different proportions of real and synthetic data. The specific configurations and mixing ratios are detailed in Section 5.2, where their impact on segmentation performance is thoroughly analyzed. Crucially, to ensure an unbiased evaluation of the model's generalization capabilities, both the validation and test sets consist exclusively of real, unmodified EM images.

As further baseline, we compare our approach against standard data augmentation techniques commonly used in biomedical imaging, including random rotations, flips (horizontal and vertical), shifts, and noise addition. Interestingly, we observed that only random rotations led to performance improvements. We attribute this to the nature of our dataset: similar to many other semantic segmentation datasets in EM [37], it is derived from a 3D volume, resulting in limited intrinsic variation. Consequently, strong augmentations can introduce unrealistic distortions that degrade performance. Through empirical testing, we found that applying random rotations within the range of $[-16°, 16°]$ yielded the best results. We refer to this configuration as "Standard Augmentations" in the remainder of the paper.

4.4. Training Details

This section outlines the specific configurations used for training both the ControlNet model and the downstream U-Net segmentation model.

ControlNet Training Our ControlNet model, built upon the Stable Diffusion 1.5 backbone, was trained for a total of 600 epochs. An Adam optimizer was used with a fixed learning rate of 1×10^{-5}. The batch size during training was set to 2. To enhance the robustness and generalization of the fine-tuning process, standard data augmentation techniques were applied to the training data. These included random cropping, the addition of random Gaussian noise, and random rotations.

U-Net Training As described above, the U-Net model, used for evaluating the quality of the generated synthetic data, was trained on combined datasets consisting of real and ControlNet-augmented images. The U-Net was trained for a maximum of 400 epochs, with an early stopping mechanism based on the convergence of the validation loss. The learning rate was set to 0.001, and an Adam optimizer was employed. A batch size of 512 was used for training the U-Net. Model performance was evaluated on the unseen test set using the mean Intersection over Union (mIoU) metric.

5. Results

5.1. Qualitative Results

Qualitative inspection of the fine-tuned ControlNet revealed that the model was able to accurately follow the spatial structure defined by the input segmentation masks (see Figure 5), effectively differentiating and visualizing the nucleus, cytoplasm, and background regions. This indicates successful alignment with the conditioning input at a structural level.

However, our analysis also uncovered notable limitations in the realism of the generated synthetic EM images, which we attribute primarily to the domain gap between ControlNet's natural image-trained backbone and the specialized characteristics of EM imagery. While the model faithfully reproduced high-frequency details along structural boundaries – such as sharp transitions between cytoplasm and background – it consistently failed to capture realistic global contrast and internal textures. In particular, it struggled to generate the intricate, high-frequency patterns within segmented regions, such as the granular textures or sub-organelle structures commonly observed in real FIB-SEM images. These internal features are essential for human observers to assess the authenticity of EM images, yet

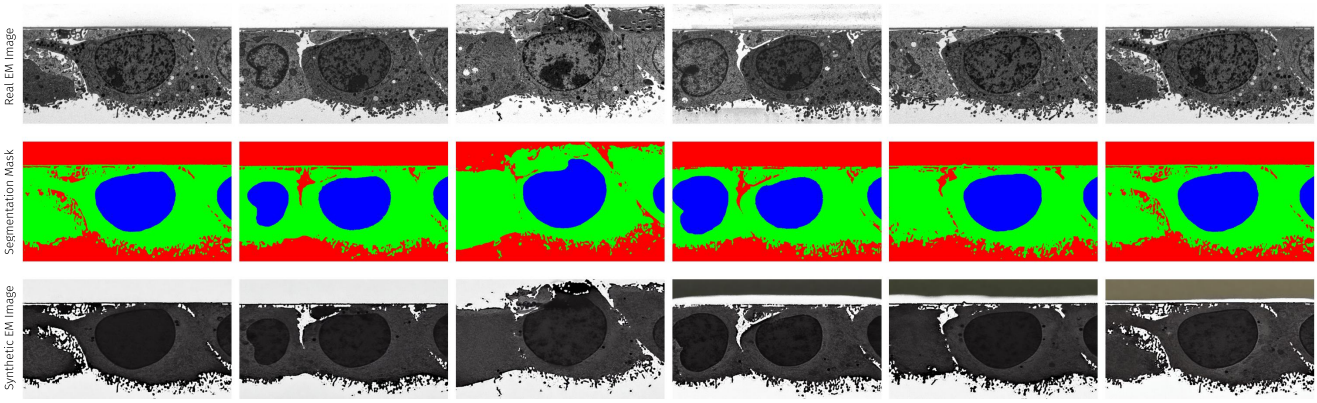

Figure 5. Representative samples from the dataset, shown alongside their corresponding color-coded segmentation masks used to condition ControlNet, and the resulting synthetic images. The visual consistency across samples (columns) highlights the low variability within the dataset. Despite visible differences between real (row 1) and synthetic (row 3) data, our quantitative experiments demonstrate that the U-Net can still extract valuable image features from the synthetic samples.

they were often missing or overly smoothed in the synthetic outputs.

We hypothesize two main factors contribute to this limitation. First, the latent-space representation used by the Stable Diffusion backbone – optimized for computational efficiency – operates in a compressed domain. During the encoding and decoding steps of the Variational Autoencoder (VAE), high-frequency information is likely suppressed or lost, making the reconstruction of fine internal textures inherently challenging. Second, because the model was pretrained on large-scale natural image datasets, it lacks prior exposure to the specialized visual features characteristic of EM data. The fine-scale structures unique to EM simply do not exist in its original training distribution.

5.2. Quantitative Results

To account for the small training data and hence the variances during the training process we train the U-Net three times and report mean and standard deviation over their resulting test mIoU.

5.2.1. Experiment 1: Pure Synthetic Data

In our first experiment, we train the U-Net exclusively on synthetically generated images (see Figure 5) and compare its performance to a model trained solely on real images. To ensure a fair comparison, both training sets are matched in size, each containing 233 samples. As expected, the model trained on synthetic data alone struggles to generalize to real images, highlighting the persistent domain gap between synthetic and real data (compare Table 1). However, despite not being exposed to any real images during training, the reported mIoU indicates that the U-Net is still able to learn and extract meaningful features from the synthetic data.

Method	Dataset Size	mIoU
Real	233	**74.6** ±0.4
ControlNet	233	44.3 ±0.4

Table 1. Comparison of U-Net performance when trained exclusively on real versus synthetically generated images. While training on synthetic data alone results in lower performance due to the domain gap, the reported mIoU indicates that the model is still able to extract meaningful features without exposure to real images.

5.2.2. Experiment 2: Dataset Augmentation

In our second experiment, we augment the training data with varying proportions of ControlNet-generated images and compare this setup to using standard augmentation techniques. While the inclusion of ControlNet-augmented data improves performance over training on real images alone, standard augmentations still outperform the ControlNet-based approach. However, as more synthetic images are added, this performance gap gradually narrows (see Figure 6). This suggests that, although ControlNet may not yet produce fully realistic EM images, the generated data is sufficiently informative to approach the dataset's saturating performance.

5.2.3. Experiment 3: Data Generation Times

In our final experiment, we assess the efficiency of data augmentation by factoring in generation time, while excluding ControlNet's training time. Specifically, we use the largest ControlNet-augmented dataset from subsubsection 5.2.2 and measure the time required to generate it. We then use this same time budget to generate standard augmentations, resulting in two datasets: 10,553 ControlNet-

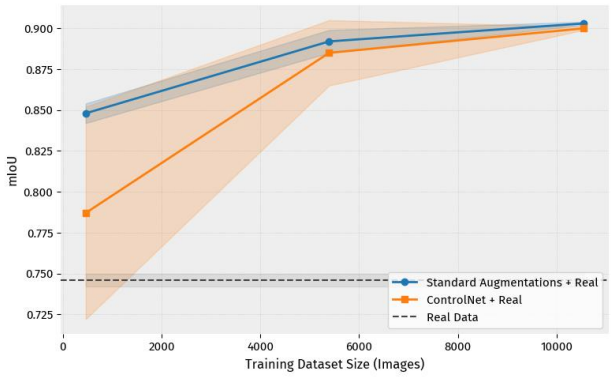

Figure 6. Comparison of segmentation performance using ControlNet-generated data and standard augmentations. ControlNet augmentations consistently outperform training on only real data. Still, while standard augmentations initially outperform ControlNet, the performance gap narrows as more synthetic data is added, approaching the dataset's saturation point.

augmented images and 34,825 standard augmented images. Consistent with the findings from Experiment 2, standard augmentations continue to outperform those generated by ControlNet Table 2, however, only by a small margin.

Method	Dataset Size	mIoU
Standard Augmentations	34,825	**90.7** $\pm$0.1
ControlNet	10,553	90.0 $\pm$0.1

Table 2. Comparison of segmentation performance using ControlNet and standard augmentations under equal generation time constraints. Despite a smaller dataset size, ControlNet-augmented data approaches the performance of standard augmentations, which still hold a slight advantage.

6. Conclusion

In this work, we successfully demonstrated the potential of fine-tuning ControlNet, based on a natural image backbone, to generate label-consistent synthetic FIB-SEM images from very small, low-variance datasets. Our synthetic data significantly boosted the performance of a downstream U-Net segmentation model, achieving results comparable to those obtained with standard augmentation techniques and yielding an impressive improvement of up to $+15.4$ mIoU compared to training without any augmentations. Importantly, standard augmentations should not be seen as competitors, but rather as complementary techniques that can be combined with diffusion-based augmentation for even greater benefit. This underlines the potential of ControlNet to enhance even low-variance datasets with domain gaps, despite its backbone being originally trained on natural images.

However, we identified limitations, particularly ControlNet's struggle to capture realistic contrast and high-frequency internal patterns within segmented regions, likely due to the inherent domain gap and the latent space compression of its pre-trained backbone. While current standard augmentations still offer a more effort-efficient path to similar performance, our study highlights ControlNet's substantial, as-yet-unrealized potential.

Future work will focus on optimizing conditioning strategies, such as combining segmentation masks with explicit edge maps, to guide the model in generating more realistic internal textures. Specifically, this refers to capturing the fine-grained structures within segmented regions (e.g., within a nucleus), which current outputs often render as overly smooth. We also believe that fine-tuning or pre-training the core generative backbone directly on large EM datasets could significantly alleviate the observed domain gap. Ultimately, our findings underscore the viability of leveraging powerful generative models to artificially enhance scarce scientific datasets, paving the way for more robust deep learning applications in domains like EM.

References

[1] CREMI: MICCAI challenge on circuit reconstruction from electron microscopy images. https://cremi.org/, 2016. Accessed July 2, 2025. 4

[2] Xiao-Chen Bai, Greg McMullan, and Sjors HW Scheres. How cryo-em is revolutionizing structural biology. *Trends in biochemical sciences*, 40(1):49–57, 2015. 1, 2

[3] Serge Beucher and Fernand Meyer. The morphological approach to segmentation: the watershed transformation. In *Mathematical morphology in image processing*, pages 433–481. CRC Press, 2018. 2

[4] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018. 3, 4

[5] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pages 205–218. Springer, 2022. 2

[6] Yue Cao, Shigang Liu, Yali Peng, and Jun Li. Denseunet: densely connected unet for electron microscopy image segmentation. *IET Image Processing*, 14(12):2682–2689, 2020. 2

[7] Vincent Casser, Kai Kang, Hanspeter Pfister, and Daniel Haehn. Fast mitochondria segmentation for connectomics. *arXiv preprint arXiv:1812.06024*, 2018. 4

[8] Winfried Denk and Heinz Horstmann. Serial blockface scanning electron microscopy to reconstruct three-dimensional tissue nanostructure. *PLoS biology*, 2(11):e329, 2004. 1

[9] Prafulla Dhariwal and Alexander Nichol. Diffusion models

beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2, 3

[10] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010. 3, 4

[11] Zehao Fang et al. Conditional diffusion model for x-ray segmentation data generation. *Journal of Artificial Intelligence Practice*, 7(1):7–10, 2024. 3

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[13] Matthew D Guay, Zeyad AS Emam, Adam B Anderson, Maria A Aronova, Irina D Pokrovskaya, Brian Storrie, and Richard D Leapman. Dense cellular segmentation for em using 2d–3d neural network ensembles. *Scientific reports*, 11 (1):2561, 2021. 4

[14] Giulio Guzzinati, Thomas Altantzis, Maria Batuk, Annick De Backer, Gunnar Lumbeeck, Vahid Samaee, Dmitry Batuk, Hosni Idrissi, Joke Hadermann, Sandra Van Aert, et al. Recent advances in transmission electron microscopy for materials science at the emat lab of the university of antwerp. *Materials*, 11(8):1304, 2018. 1

[15] Yang Heng, Ma Yinghua, Fiaz Gul Khan, Ahmad Khan, Farman Ali, Ahmad Ali AlZubi, and Zeng Hui. Survey: application and analysis of generative adversarial networks in medical images. *Artificial Intelligence Review*, 58(2):39, 2024. 2

[16] David Grant Colburn Hildebrand, Marcelo Cicconet, Russel Miguel Torres, Woohyuk Choi, Tran Minh Quan, Jungmin Moon, Arthur Willis Wetzel, Andrew Scott Champion, Brett Jesse Graham, Owen Randlett, et al. Whole-brain serial-section electron microscopy in larval zebrafish. *Nature*, 545(7654):345–349, 2017. 1, 3

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 3

[18] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33, 2022. 2, 3

[19] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 2

[20] Michael Kass, Andrew Witkin, and Demetri Terzopoulos. Snakes: Active contour models. *International journal of computer vision*, 1(4):321–331, 1988. 2

[21] Abid Khan, Chia-Hao Lee, Pinshane Y Huang, and Bryan K Clark. Leveraging generative adversarial networks to create realistic scanning transmission electron microscopy images. *npj Computational Materials*, 9(1):85, 2023. 2

[22] C Kizilyaprak, J Daraspe, and BM Humbel. Focused ion beam scanning electron microscopy in biology. *Journal of microscopy*, 254(3):109–114, 2014. 1

[23] Graham Knott, Herschel Marchman, David Wall, and Ben Lich. Serial section scanning electron microscopy of adult brain tissue using focused ion beam milling. *Journal of Neuroscience*, 28(12):2959–2964, 2008. 1

[24] Mingxing Li, Chang Chen, Xiaoyu Liu, Wei Huang, Yueyi Zhang, and Zhiwei Xiong. Advanced deep networks for 3d mitochondria instance segmentation. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2022. 2

[25] Ziyi Li, Qinye Zhou, Xiaoyun Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Open-vocabulary object segmentation with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7667–7676, 2023. 2, 3

[26] Weibin Liao, Yinghao Zhu, Xinyuan Wang, Chengwei Pan, Yasha Wang, and Liantao Ma. Lightm-unet: Mamba assists in lightweight unet for medical image segmentation. *arXiv preprint arXiv:2403.05246*, 2024. 2

[27] Chixiang Lu, Kai Chen, Heng Qiu, Xiaojun Chen, Gu Chen, Xiaojuan Qi, and Haibo Jiang. Emdiffuse: a diffusion-based deep learning method augmenting ultrastructural imaging and volume electron microscopy. *bioRxiv*, pages 2023–07, 2023. 3

[28] Nan Nan and Jingxin Wang. Fib-sem three-dimensional tomography for characterization of carbon-based materials. *Advances in Materials Science and Engineering*, 2019(1): 8680715, 2019. 1

[29] Christopher J Peddie, Christel Genoud, Anna Kreshuk, Kimberly Meechan, Kristina D Micheva, Kedar Narayan, Constantin Pape, Robert G Parton, Nicole L Schieber, Yannick Schwab, et al. Volume electron microscopy. *Nature Reviews Methods Primers*, 2(1):51, 2022. 1, 2

[30] Torben Prill, Katja Schladitz, Dominique Jeulin, Matthieu Faessel, and C Wieser. Morphological segmentation of fib-sem data of highly porous media. *Journal of microscopy*, 250(2):77–87, 2013. 1

[31] Jean-Paul Renaud, Ashwin Chari, Claudio Ciferri, Wen-ti Liu, Hervé-William Rémigy, Holger Stark, and Christian Wiesmann. Cryo-em in drug discovery: achievements, limitations and prospects. *Nature reviews Drug discovery*, 17(7): 471–492, 2018. 1, 2

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 3, 5

[33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 2, 6

[34] Jiacheng Ruan, Mingye Xie, Jingsheng Gao, Ting Liu, and Yuzhuo Fu. Ege-unet: an efficient group enhanced unet for skin lesion segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 481–490. Springer, 2023. 2

[35] Martin Salzer, Aaron Spettl, Ole Stenzel, Jan-Henrik Smått, Mika Lindén, Ingo Manke, and Volker Schmidt. A two-stage approach to the segmentation of fib-sem images of highly porous materials. *Materials Characterization*, 69:115–126, 2012. 1

[36] Kavitha Shaga Devan, Paul Walther, Jens von Einem, Timo Ropinski, Hans A Kestler, and Clarissa Read. Improved automatic detection of herpesvirus secondary envelopment stages in electron microscopy by augmenting training data with synthetic labelled images generated by a generative adversarial network. *Cellular Microbiology*, 23(2):e13280, 2021. 2, 3

[37] Kavitha Shaga Devan, Hans A Kestler, Clarissa Read, and Paul Walther. Weighted average ensemble-based semantic segmentation in biological electron microscopy images. *Histochemistry and Cell Biology*, 158(5):447–462, 2022. 1, 2, 3, 4, 6

[38] A Sheidaei, M Baniassadi, M Banu, P Askeland, M Pahlavanpour, Nick Kuuttila, F Pourboghrat, LT Drzal, and H Garmestani. 3-d microstructure reconstruction of polymer nano-composite using fib–sem and statistical correlation function. *Composites Science and Technology*, 80:47–54, 2013. 1

[39] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019. 2, 3

[40] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019. 2, 3

[41] Nima Tajbakhsh, Jae Y Shin, Suryakanth R Gurudu, R Todd Hurst, Christopher B Kendall, Michael B Gotway, and Jianming Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312, 2016. 1, 3

[42] Abhishek Thakur and Gopal Kumar Thakur. Developing gans for synthetic medical imaging data: Enhancing training and research. *Int. J. Adv. Multidiscip. Res*, 11(1):70–82, 2024. 2

[43] Clarissa Villinger, Heiko Gregorius, Christine Kranz, Katharina Höhn, Christin Münzberg, Götz von Wichert, Boris Mizaikoff, Gerhard Wanner, and Paul Walther. Fib/sem tomography with tem-like resolution for 3d imaging of high-pressure frozen cells. *Histochemistry and cell biology*, 138: 549–556, 2012. 5

[44] Jason Wang, Luis Perez, et al. The effectiveness of data augmentation in image classification using deep learning. *Convolutional Neural Networks Vis. Recognit*, 11(2017):1–8, 2017. 2, 3

[45] Junde Wu, Rao Fu, Huihui Fang, Yu Zhang, Yehui Yang, Haoyi Xiong, Huiying Liu, and Yanwu Xu. Medsegdiff: Medical image segmentation with diffusion probabilistic model. In *Medical Imaging with Deep Learning*, pages 1623–1639. PMLR, 2024. 2, 3

[46] Hanrong Ye, Jason Kuen, Qing Liu, Zhe Lin, Brian Price, and Dan Xu. Seggen: Supercharging segmentation models with text2mask and mask2img synthesis. In *European Conference on Computer Vision*, pages 352–370. Springer, 2024. 2, 3

[47] Boxiao Yu and Kuang Gong. Adaptive whole-body pet image denoising using 3d diffusion models with controlnet. *arXiv preprint arXiv:2411.05302*, 2024. 3

[48] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 2, 3, 5

[49] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 3, 4