Joint EM Image Super-Resolution and Segmentation with Semantic and Structural Priors

Jiateng Shou[1,2] **Haoyuan Shi**[1,2] **Zhiwei Xiong**[1,2*]
[1]MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition,
University of Science and Technology of China
[2]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
shoujt@mail.ustc.edu.cn zwxiong@ustc.edu.cn

Abstract

Reconstructing the connectome of the brain is a central goal of modern neuroscience, but the required electron microscopy (EM) imaging is extremely time-consuming. EM image super-resolution (SR) accelerates this acquisition for downstream neuron segmentation, yet existing methods do not adequately couple SR with segmentation for EM: SR lacks segmentation guidance, and segmentation cannot adapt to super-resolved features. In this work, we propose a joint SR and segmentation framework specifically designed for EM images. In the super-resolution part, we introduce segmentation as an auxiliary task via a shared encoder with task-specific decoders to provide a semantic prior, improving reconstruction fidelity. In the segmentation part, we introduce a lightweight partial diffusion model that initializes denoising from a partially noised SR embedding, producing a structural prior that closely approximates real high-resolution features and enhances segmentation accuracy on super-resolved EM images. Extensive experiments across multiple datasets and degradation settings demonstrate the superior performance of our framework. Code is available at https://github.com/jtshou/EMSRSeg.

1 Introduction

Reconstructing the connectome—the synaptic-resolution wiring diagram of the brain—is a central goal of modern neuroscience [1, 27]. Only electron microscopy (EM) resolves individual axons and dendrites at the nanometer scale required for circuit reconstruction, but acquiring EM data at meaningful spatial extent takes months to years [32]. To translate the resulting volumes into circuit, neuron segmentation must delineate individual neurons for downstream connectivity analysis.

To accelerate imaging, deep learning-based EM image super-resolution (EMSR) [9, 34, 24] has been increasingly developed to reconstruct high-resolution (HR) images from fast, low-resolution (LR) scans. However, existing EMSR methods are optimized in isolation, focusing on pixel-level reconstruction while ignoring their interaction with downstream segmentation. In practice, super-resolved EM images are directly used as inputs for segmentation, yet the lack of task-aware design leads to suboptimal segmentation performance. While joint SR-segmentation frameworks have been explored in natural and medical imaging [36, 29, 15], existing frameworks fall short of the bidirectional task coupling that EM requires: SR reconstruction needs to benefit from segmentation guidance, while segmentation needs to remain robust to the feature gap introduced by SR.

As illustrated in Figure 1, existing joint frameworks can be roughly categorized into two groups: architectures with a shared encoder and task-specific decoders, and cascaded SR–segmentation framework. Shared-encoder architectures [36, 35, 29] typically treat segmentation as the primary task and attempt to learn a shared representation through feature constraints or auxiliary objectives.

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

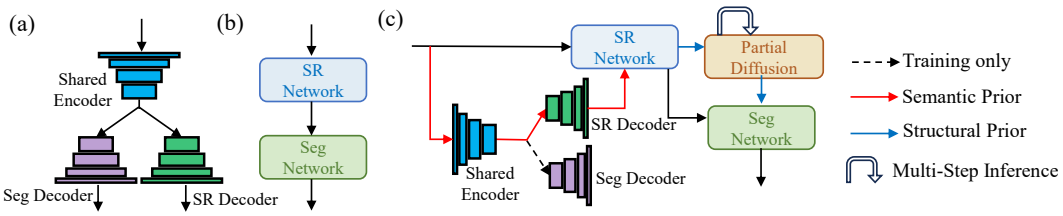

Figure 1: Comparison of existing frameworks with our framework. (a) Architecture with a shared encoder and task-specific decoders. (b) Cascaded SR–segmentation framework. (c) Our framework.

However, such frameworks perform poorly on EM data due to the structural mismatch between SR and EM segmentation networks and the difficulty of accurately upsampling affinity-based predictions.

Cascaded frameworks can be further divided into two types. Perceptual-oriented frameworks first train advanced SR networks [34, 42] and segmentation networks [18, 22] independently, applying SR to ensure visual fidelity before performing segmentation. However, the distribution gap between super-resolved and real HR EM images often leads to suboptimal segmentation. Segmentation-oriented frameworks [15, 13] design SR models that favor segmentation accuracy, usually at the cost of reconstruction quality. Overall, cascaded frameworks are limited by their decoupled design: SR networks do not leverage semantic guidance from segmentation, and segmentation networks do not account for discrepancies between super-resolved and real HR features.

In contrast, we propose a new framework that explicitly couples SR and segmentation through semantic and structural priors. In the SR part, we introduce an auxiliary segmentation branch during training, using a shared encoder to provide a semantic prior. This prior is injected into the SR network through an attention mechanism, preserving the original SR backbone while enhancing structural consistency. To exploit spatial correlations across EM images, we further propose a Block-wise Window-based Alignment (BWA) module that aligns EM features within local windows in a single step, providing cross-section feature aggregation tailored to 3D EM volumes.

In the segmentation part, we incorporate a structural prior from a lightweight partial diffusion model to reduce the feature distribution gap between super-resolved images and real HR EM images. The structural prior is intended to approximate the prior extracted from real HR EM images, and we adopt a two-stage training strategy. In Stage I, an HR prior extracted from real HR EM images is directly injected into the segmentation part and jointly optimized with the SR network. In Stage II, we freeze the trained SR module and fine-tune the segmentation part, replacing the HR prior with a structural prior from a partial diffusion model learned to approximate it. By initializing denoising from a partially noised SR embedding rather than Gaussian noise, the trajectory stays closer to the HR distribution throughout, producing a more accurate prior.

Experimental results demonstrate that, unlike previous frameworks that trade SR fidelity for segmentation gains, our framework improves *both*: it surpasses the standalone SR baseline in reconstruction quality, while the structural prior consistently boosts segmentation over the no-prior counterpart.

Our main contributions are as follows: (1) We propose the first unified framework for EM image super-resolution and segmentation. (2) For SR, we introduce an auxiliary segmentation branch to provide a semantic prior, and design a BWA module to exploit cross-section structural continuity, improving SR performance. (3) For segmentation, we introduce a lightweight partial diffusion model that denoises from a partially noised SR embedding, producing a structural prior that closely approximates real HR features and improves segmentation accuracy. (4) Extensive experiments across multiple datasets and degradation settings demonstrate the effectiveness of our framework.

2 Related work

2.1 EM Image Super-Resolution

Deep learning–based EMSR methods can be categorized according to whether they exploit correlations across serial EM sections. Single-image EMSR methods [7, 28, 41, 24] operate on individual sections and thus cannot leverage the structural continuity inherent in 3D EM volumes. Multi-section

methods address this by aggregating neighboring slices: PSSR [9] reconstructs the center section with a ResNet–UNet, and GPEMSR [34] further integrates generative priors for perceptual quality. More recently, transformer-based SR models [20, 33, 42] have demonstrated strong performance in terms of reconstruction fidelity. However, existing SR methods largely overlook the semantic priors in EM images, limiting recovery of fine cellular structures under severe degradation.

2.2 EM Image Segmentation

Unlike natural image instance segmentation, EM segmentation typically relies on an affinity-based paradigm that predicts three-channel voxel affinities and recovers instances through non-differentiable post-processing [11, 3]. Building upon this formulation, Superhuman [18] adopts an anisotropic UNet and achieves strong performance in neuron instance segmentation, while MALA [11] further improves accuracy by scaling to a larger 3D UNet and introducing the MALIS loss. To improve efficiency and alleviate block-boundary artifacts in large-volume inference, CAD [22] distills priors from 3D Superhuman network to 2D segmentation network. More recently, EMmamba [6] proposes a pretraining strategy for Mamba-based architectures. However, these methods assume real HR inputs and overlook the structural degradation introduced by SR.

2.3 Joint SR–Segmentation Framework

Existing joint frameworks generally fall into two categories: architectures with a shared encoder and task-specific decoders, and cascaded SR–segmentation framework. Shared-encoder architectures [36, 35, 29] typically impose constraints on shared features within a multi-task learning framework, where segmentation is often treated as the primary task. Cascaded frameworks, on the other hand, decouple SR and segmentation into sequential stages, either by directly applying segmentation on super-resolved images or by designing SR models tailored for segmentation-aware reconstruction [15, 13]. Recently, diffusion-based cascaded frameworks [5, 16] have shown promising performance in both reconstruction and segmentation; however, these methods typically rely on Stable Diffusion [31] pretrained on natural images, which are not suited to EM data. In contrast, we propose a novel joint SR and segmentation framework specifically designed for EM images, achieving improved performance in both SR and segmentation.

3 Method

3.1 Overview

As illustrated in Figure 2, given a sequence of low-resolution EM images $\{I^z_{\text{LR}} \in \mathbb{R}^{H\times W\times 1} \mid z = 1, 2, \ldots, Z\}$, our goal is to reconstruct high-resolution images $I^z_{\text{SR}} \in \mathbb{R}^{rH\times rW\times 1}$ that closely approximate their ground-truth counterparts I^z_{GT}, while simultaneously producing high-resolution segmentation maps Seg^z_{SR} that match Seg^z_{GT}. Here, r denotes the super-resolution scale factor.

To accomplish this, we introduce a two-stage learning framework. Stage I performs semantic-prior-guided super-resolution and HR-prior-guided segmentation. Specifically, in the SR part, we treat segmentation as an auxiliary task and inject a semantic prior from the shared encoder to guide the SR network, yielding the preliminary super-resolved images I^z_{SR}. Then, we further introduce an HR prior from the ground-truth HR images into the segmentation backbone to obtain the corresponding affinity maps A^z_{HR}. Following common practice in EM segmentation [11, 22], the network is supervised on voxel-wise affinities rather than explicit instance labels. An affinity measures the probability that two neighboring voxels (e.g., along the h, w, or z direction) belong to the same instance.

Stage II performs structural-prior-guided segmentation. Specifically, we freeze the learned SR part and remove the segmentation decoder. We then train a partial diffusion model while fine-tuning the segmentation backbone. The structural prior derived from a partial diffusion model is injected into the segmentation backbone to replace the HR prior, producing affinity maps A^z_{SR}. The final segmentation Seg^z_{SR} is then obtained via the Waterz post-processing algorithm [11].

3.2 Semantic-Prior-Guided Super-Resolution

High-level semantic priors have proven beneficial for image restoration [43, 40, 37], yet transferring them to EMSR is non-trivial. One line of work distills priors from natural-image segmentation

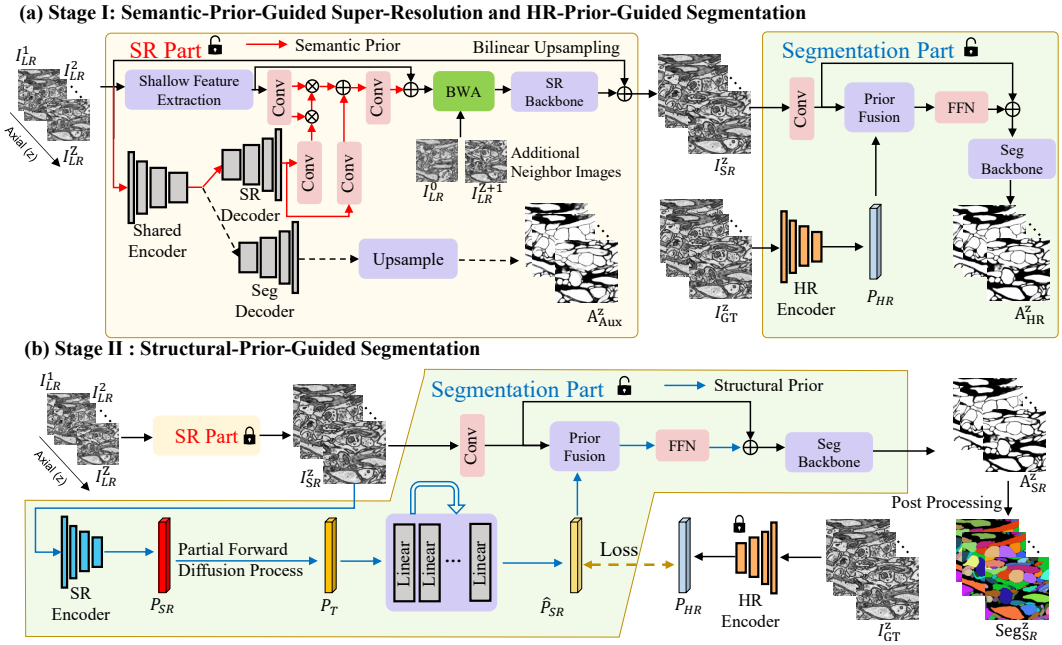

Figure 2: Overview of our framework. Stage I introduces a semantic prior for SR and an HR image prior for segmentation through joint training. Stage II replaces the HR prior with a structural prior from a partial diffusion model while fixing the SR part.

networks [40, 43], but a substantial domain gap in both visual appearance and semantic categories prevents effective transfer to EM. The other shares an encoder between the SR and segmentation branches [29, 30], but has not been adapted to EM, where two domain-specific obstacles arise. First, SR backbones [4, 44, 19] and EM segmentation backbones [18, 22] differ substantially, so naively sharing an encoder yields ineffective joint representations. Second, EM segmentation predicts voxel affinities rather than semantic masks, and an affinity decoder operating on upsampled SR features tends to produce blurry, structurally discontinuous boundaries.

Based on these observations, we retain the standard SR architecture and use the shared-encoder features only as a semantic prior fused into the SR part. Since the segmentation decoder alone does not yield reliable affinity predictions, we employ this decoder solely during training to learn the priors and discard it at inference time. The SR decoder output is fused with shallow features extracted from the early layers of the SR network through a cross-attention mechanism. Specifically, let F^z_{shallow} denote the shallow features and F^z_{dec} denote the SR decoder output. We first obtain an attention map

$$A^z = \sigma\big(\text{Conv}(F^z_{\text{shallow}}) \otimes \text{Conv}(F^z_{\text{dec}})\big), \tag{1}$$

where $\sigma(\cdot)$ is the sigmoid function and $\otimes$ denotes element-wise multiplication. $\text{Conv}(\cdot)$ denotes a convolution operation. The final fused feature is computed as

$$F^z_{\text{fusion}} = \text{Conv}(A^z \otimes \text{Conv}(F^z_{\text{shallow}}) + \text{Conv}(F^z_{\text{dec}})) + F^z_{\text{shallow}}. \tag{2}$$

This fusion injects a semantic prior into the SR process while maintaining backbone compatibility.

Adjacent EM sections exhibit strong structural continuity, and leveraging this cross-section information is crucial for improving SR quality. To more effectively utilize these inter-section dependencies, we introduce a Block-wise Window-based Alignment (BWA) module that aligns EM features across neighboring sections. As illustrated in Figure 3, unlike the second-order alignment-and-propagation strategy commonly used in video SR [4], our strategy requires only a single alignment step. For each super-resolved 3D EM volume, referred to as a *block*, we first partition the block into multiple overlapping windows. Within each window, features from all sections are aligned toward the center section using deformable convolution [46]. To avoid zero-padding or other artificial boundary handling, we incorporate the nearest out-of-block sections, denoted as I^0_{LR} and I^{Z+1}_{LR}. Their shallow features and self-attention-enhanced representations are used to strengthen the SR result of the head

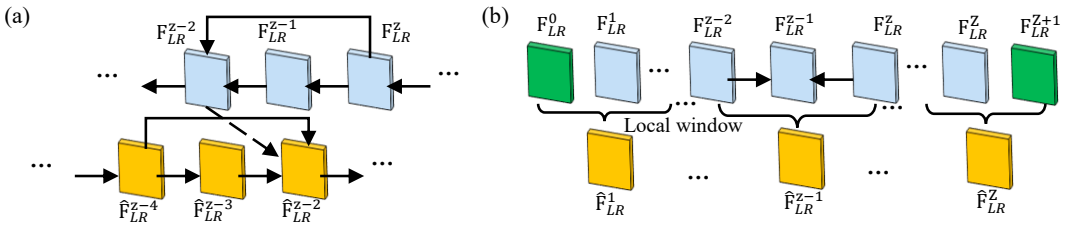

Figure 3: Comparison of existing alignment strategy with our proposed BWA module. (a) Multi-image alignment strategy in video super-resolution. (b) Our alignment strategy, i.e., the BWA module.

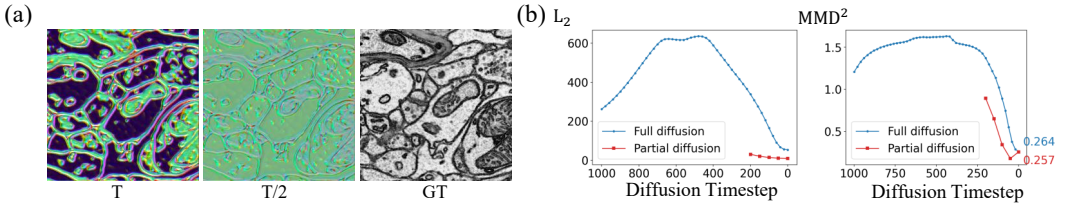

Figure 4: Analysis of partial vs. full diffusion priors. (a) Visualization of the Prior Fusion output when the noised prior at $t=T$ (start) and $t=T/2$ (mid-point) is fed into the fusion module: the noised SR prior already encodes coarse cellular structure at $t=T$, and denoising progressively suppresses organelle-interior activations while preserving neuron-instance boundaries. (b) Distance between the prior trajectories and the GT noisy distribution $q(z_t \mid P_{HR})$, measured by mean L_2 distance and unbiased MMD^2 [12], averaged over the Wafer 4 test set. **Lower** is closer to the GT distribution.

and tail sections of the block. For each section, the aligned features are fused within a local window to produce the corresponding output $\hat{F}^z_{\mathrm{LR}}$. In total, the BWA module adds only 0.31M parameters.

3.3 Structural-Prior-Guided Segmentation

A key challenge for segmentation on super-resolved EM images is the feature distribution gap from real HR images. Diffusion-based SR [38, 39] can narrow this gap by restoring missing details, but sacrifices reconstruction fidelity and incurs heavy sampling. We instead bring diffusion into the segmentation part as a lightweight structural prior: a compact diffusion model operates on a per-section embedding rather than on pixels, and is trained to reproduce the prior extracted from real HR EM images, yielding HR-aligned guidance without altering the SR output.

We adopt a two-stage strategy to learn the structural prior. In Stage I, we extract a target HR prior P_{HR} from real HR EM images via an HR encoder and inject it into the segmentation branch through a Prior Fusion process (see Appendix A for details), followed by a feed-forward network. The enriched features are then processed by the segmentation backbone to produce the affinity map A^z_{HR}. We jointly train the HR-prior-guided segmentation and the super-resolution network in Stage I.

In Stage II, since real HR EM images are unavailable at inference, we learn a diffusion model that produces a structural prior $\hat{P}_{\mathrm{SR}}$ approximating P_{HR}. Inspired by [21], we compress each EM section into a 1×1 vector so that the diffusion backbone reduces to a few linear layers, adding only 1.63M parameters in total. While [21] denoises from pure Gaussian noise, our key insight, illustrated in Figure 4, is that the SR-side embedding already carries substantial structural information, so initializing the trajectory from a partially noised SR embedding yields noisy-distribution estimates that more closely match the GT noisy distribution $q(z_t \mid P_{HR})$ at every timestep, leading to more accurate guidance with far fewer denoising steps.

We support this insight with two observations. As shown in Figure 4(a), we feed diffusion features at $t=T$ and $t=T/2$ into the Prior Fusion process and visualize the predicted instance boundaries: even at $t=T$ the prior activates on both organelle interiors and neuron boundaries, and denoising progressively suppresses the interior activations, a trend already visible at $t=T/2$, where organelle interiors are largely treated as background. Figure 4(b) further plots the L_2 distance and unbiased MMD^2 [12] between the prior trajectory and the GT noisy distribution $q(z_t \mid P_{\mathrm{HR}})$: the full-diffusion

curve exhibits a pronounced "hump" at mid-range timesteps ($t \approx 400$-700), where noise overwhelms the signal and the prior drifts farthest from the GT noisy distribution. We therefore initialize diffusion at a partial timestep $T = 200$ that lies strictly within the signal-dominated regime, bypassing the hump entirely.

Formally, let P_{SR} denote the high-dimensional embedding of an EM section obtained from the encoder. The noising process is defined as

$$P_T = \sqrt{\bar{\alpha}_T}\, P_{\mathrm{SR}} + \sqrt{1 - \bar{\alpha}_T}\, \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), \tag{3}$$

where T is the partial starting timestep and $\bar{\alpha}_t$ follows the standard definition in [14].

With this starting point, a few denoising steps suffice; neither more steps nor a full process from Gaussian noise further improves the prior. For an n-step inference schedule, we select a discrete set of timesteps following [25] $\{t_0 = T,\, t_1 = \lfloor T(n-1)/n \rfloor,\, t_2 = \lfloor T(n-2)/n \rfloor,\, \ldots,\, t_{n-1} = \lfloor T/n \rfloor\}$. Given the starting point $\hat{P}_{t_0} = P_T$, the inference process proceeds as

$$\hat{P}_{t_{k+1}} = \frac{\sqrt{\bar{\alpha}_{t_{k+1}}}\beta_{t_k}}{1 - \bar{\alpha}_{t_k}} D_\theta(\hat{P}_{t_k}, t_k) + \frac{\sqrt{\alpha_{t_k}}(1 - \bar{\alpha}_{t_{k+1}})}{1 - \bar{\alpha}_{t_k}} \hat{P}_{t_k} + \sqrt{\beta_{t_k}}\epsilon, \alpha_{t_k} = \frac{\bar{\alpha}_{t_k}}{\bar{\alpha}_{t_{k+1}}}, \beta_{t_k} = 1 - \alpha_{t_k}, \tag{4}$$

where $k = 0, 1, \ldots, n-2$, and D_θ denotes the denoised the diffusion backbone. In the final step, no noise is added, and the output is denoted as $\hat{P}_{\mathrm{SR}}$,

$$\hat{P}_{\mathrm{SR}} = D_\theta(\hat{P}_{t_{n-1}}, t_{n-1}). \tag{5}$$

3.4 Training Strategy

In Stage I, the overall training objective consists of three loss terms, corresponding to super-resolution, auxiliary segmentation, and HR-prior-guided segmentation.

$$\mathcal{L}_{\mathrm{S1}} = \left\| I^z_{\mathrm{SR}} - I^z_{\mathrm{GT}} \right\|_1 + \left\| W \otimes (A^z_{\mathrm{Aux}} - A^z_{\mathrm{GT}}) \right\|_2^2 + \left\| W \otimes (A^z_{\mathrm{HR}} - A^z_{\mathrm{GT}}) \right\|_2^2, \tag{6}$$

where W is a per-voxel weight map computed by inverse-frequency rebalancing on the affinity map to upweight sparse boundary voxels, following CAD [22].

Stage II freezes the SR module and the HR encoder, locking P_{HR} as the target distribution. A diffusion model is then trained to produce $\hat{P}_{\mathrm{SR}}$ that approximates P_{HR} from super-resolved inputs

$$\mathcal{L}_{\mathrm{S2}} = \left\| W \otimes (A^z_{\mathrm{SR}} - A^z_{\mathrm{GT}}) \right\|_2^2 + \lambda_{\mathrm{diff}} \left\| \hat{P}_{\mathrm{SR}} - P_{\mathrm{HR}} \right\|_2^2, \tag{7}$$

where the second term distinguishes our setup from standard DDPM training: the target P_{HR} is itself a learned, task-shaped representation rather than a fixed noise schedule.

4 Experiments

4.1 Settings

Datasets. We conduct experiments on two EM datasets to evaluate the performance of our framework on both super-resolution and segmentation tasks. The first dataset is MEC dataset [6], which contains six volumes acquired from multiple regions of the mouse brain (wafer 4/25/26/26-2/36/36-2), each with a resolution of $1250 \times 1250 \times 125$ voxels and dense segmentation annotations, making it one of the largest publicly available EM datasets with voxel-level ground truth. We center-crop each slice to 1024×1024, and follow [6] to split the dataset into training (wafer 25/26/26-2/36), validation (wafer 4), and testing (wafer 36-2) sets. We evaluate $8\times$ bicubic downsampling and a $4\times$ setting that sequentially applies Gaussian beam blur with $\sigma = 1.5$ HR pixels, bilinear downsampling, and Poisson shot noise with a peak expected count of 60. The second dataset is CREMI [10], which consists of three sub-volumes (A, B, and C) from the adult *Drosophila melanogaster* brain, each of size $1250 \times 1250 \times 125$ voxels. We select CREMI for its high-quality imaging and similarly center-crop each slice to 1024×1024, following [22] by using the first 100 slices for training and the remaining 25 slices for testing. To demonstrate robustness under different degradation settings, we apply $8\times$ bicubic downsampling to CREMI A, $8\times$ bicubic downsampling with additive Gaussian noise ($\sigma \in [10, 25]$) to CREMI B, and a more challenging $16\times$ bicubic downsampling to CREMI C.

Table 1: Quantitative comparison of visual quality and EM image segmentation results on the MEC dataset. "+" indicates that the super-resolution and segmentation networks are jointly optimized. The best and the second-best results are highlighted in **bold** and underline.

Methods		Wafer 4				Wafer 36-2				Params/M
Super-resolution	Segmentation	PSNR↑	SSIM↑	VOI↓	ARAND↓	PSNR↑	SSIM↑	VOI↓	ARAND↓	
BasicVSR++	Superhuman	17.60	0.4221	4.2943	0.8176	17.65	0.4900	2.9699	0.3750	9.02
	CAD			4.1748	0.7480			3.0646	0.4751	8.37
GPEMSR	Superhuman	17.72	0.4305	3.4248	0.7475	17.95	0.5103	2.0101	0.2751	70.80
	CAD			3.1837	0.7121			1.3728	0.2173	70.14
OFTSR	Superhuman	17.61	0.4290	3.8709	0.7770	17.70	0.5005	2.5658	0.3443	95.11
	CAD			3.7804	0.7284			2.2755	0.3436	94.46
IART	Superhuman	17.76	0.4359	3.9646	0.7722	17.87	0.5097	2.5297	0.3325	14.06
	CAD			3.9973	0.7470			2.2081	0.3144	13.40
IART + Superhuman		17.19	0.3914	1.5356	0.1629	17.24	0.4557	0.9756	0.2074	14.02
DS2F		16.41	0.3798	3.2944	0.3471	16.49	0.4283	2.9904	0.5692	20.81
SR4IR		17.21	0.4148	1.6064	0.3293	17.18	0.4785	0.7701	0.1114	13.52
Ours		**17.99**	**0.4466**	**1.3726**	**0.1230**	**18.28**	**0.5327**	**0.5971**	**0.0934**	8.65

Table 2: Quantitative comparison of visual quality and EM image segmentation results on the CREMI dataset. "+" indicates that the super-resolution and segmentation networks are jointly optimized. The best and the second-best results are highlighted in **bold** and underline.

Methods		CREMI A SR (8×)				CREMI B SR (8×) + noise ($\sigma \in [10, 25]$)				CREMI C SR (16×)			
Super-resolution	Segmentation	PSNR↑	SSIM↑	VOI↓	ARAND↓	PSNR↑	SSIM↑	VOI↓	ARAND ↓	PSNR↑	SSIM↑	VOI↓	ARAND↓
BasicVSR++	Superhuman	30.77	0.8237	0.7741	0.0915	24.93	0.6671	3.0169	0.4729	22.82	0.5591	3.2039	0.3287
	CAD			0.7670	0.1165			2.7715	0.2746			4.0470	0.4630
GPEMSR	Superhuman	30.89	0.8225	0.7582	0.0909	24.66	0.6512	2.6800	0.1920	22.75	0.5505	3.0550	0.2824
	CAD			0.7560	0.1123			2.9752	0.2940			3.9181	0.4378
IART	Superhuman	31.03	0.8281	0.7595	0.0903	24.88	0.6572	4.0798	0.6654	22.62	0.5491	3.8690	0.4108
	CAD			0.7516	0.1114			3.3360	0.3761			5.0567	0.6191
IART + Superhuman		30.20	0.8063	0.7152	0.1073	24.19	0.5945	1.8364	0.1660	22.20	0.4918	3.0379	0.2850
DS2F		25.10	0.5821	1.8280	0.4646	23.46	0.4999	3.0766	0.2954	21.21	0.5079	5.4758	0.7306
SR4IR		30.52	0.8154	0.7270	0.1126	24.32	0.6164	1.7061	0.1610	21.89	0.4777	2.7879	0.2509
Ours		**31.13**	**0.8318**	**0.6717**	**0.0893**	**25.13**	**0.6724**	**1.6332**	**0.1038**	**23.01**	**0.5645**	**2.4881**	**0.2083**

Metrics. To evaluate SR performance, we measure fidelity using PSNR and SSIM. To evaluate segmentation performance, we adopt two commonly used metrics in EM segmentation: VOI [2] and ARAND [26]. These metrics assess the differences from the ground truth from different perspectives.

Backbone selection. For the SR part, we adopt different backbones on different datasets: a CNN-based backbone on the MEC dataset and a Transformer-based backbone on the CREMI dataset, to demonstrate that our framework is effective across different architectural choices. For the CNN backbone, we employ the 3DA module proposed in [34] to fuse the aligned features in the BWA module, while for the Transformer backbone, we use a 3D Swin Transformer [33] for feature fusion. For the segmentation part, we consistently adopt the widely used Superhuman backbone [18]. Full comparisons of both SR backbones are provided in Appendix E.

Training setting. We train the framework using the Adam optimizer [17] ($\beta_1 = 0.9$, $\beta_2 = 0.99$) with a cosine annealing learning rate schedule [23]. The initial learning rate is set to 1×10^{-4}. On the MEC dataset and the CREMI A dataset, we set a higher initial learning rate of 1×10^{-3} for the segmentation decoder and the segmentation part in Stage I, while keeping the remaining components unchanged, which empirically leads to better performance. All models are trained on two NVIDIA A100 GPUs. The diffusion model is defined with a linear noise schedule. The number of diffusion steps is set to 4, as justified by the ablation study. We set $Z = 9$, corresponding to 9 input sections per sample, and the ground-truth resolution during training is fixed to 256×256. We further set $\lambda_{\text{diff}} = 0.1$ so that the two terms in L_{S2} are on a comparable scale. For post-processing,Waterz merges adjacent supervoxels when the median affinity across their boundary exceeds 0.5, a single threshold used for all methods following CAD [22].

4.2 Comparisons with Existing Methods

We compare our framework with three categories of methods. (1) Perceptual-oriented cascaded frameworks, which combine advanced video SR methods (BasicVSR++ [4], IART [42]), diffusion SR method (OFTSR [47]) or EMSR method (GPEMSR [34]) with advanced EM segmentation methods (Superhuman [18], CAD [22]). In these pipelines, the SR and segmentation models are trained separately, where the SR model is optimized using LR–HR image pairs, and the segmentation model

Table 3: Quantitative comparison on the MEC dataset at $4\times$ with beam blur and shot noise. "+" indicates that the super-resolution and segmentation networks are jointly optimized. The best and the second-best results are highlighted in **bold** and underline.

Methods		Wafer 4				Wafer 36-2			
Super-resolution	Segmentation	PSNR↑	SSIM↑	VOI↓	ARAND↓	PSNR↑	SSIM↑	VOI↓	ARAND↓
IART	Superhuman	<u>18.90</u>	0.5341	2.6477	0.4827	<u>19.08</u>	<u>0.5995</u>	1.0328	0.1627
	CAD			3.0134	0.5370			1.3884	0.2470
IART + Superhuman		18.61	0.5216	1.8770	0.4895	18.79	0.5803	0.7633	0.1510
SR4IR		18.78	<u>0.5349</u>	<u>1.8360</u>	<u>0.4562</u>	19.00	0.5934	<u>0.6224</u>	<u>0.1061</u>
Ours		**19.10**	**0.5439**	**1.6280**	**0.3251**	**19.50**	**0.6226**	**0.5572**	**0.0971**

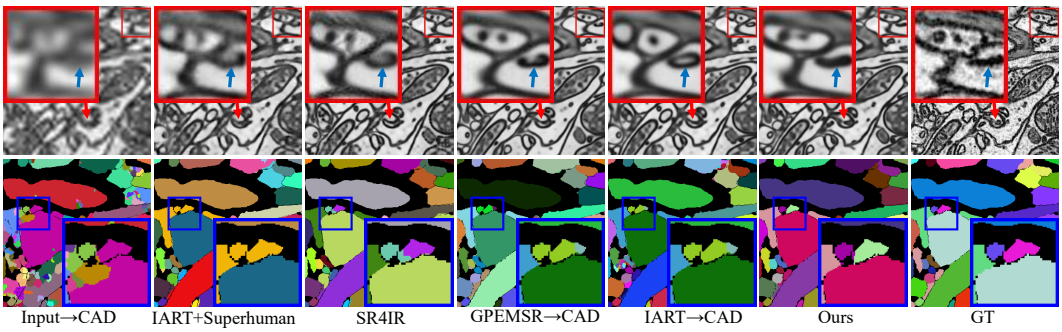

Figure 5: Qualitative comparison of $8\times$ bicubic-downsampled super-resolution and segmentation results on the wafer 4 dataset. "+" indicates joint optimization of SR and segmentation, whereas "→" denotes a cascaded framework with separate training. The input is obtained by bicubic interpolation of the LR image. We select methods with the best quantitative performance for comparison.

is trained using HR images and their corresponding segmentation labels. (2) Segmentation-oriented cascaded frameworks, where we cascade the high-performing SR method IART with Superhuman and jointly optimize the entire pipeline. We also compare with SR4IR [15], where the original 2D natural-image segmentation network is replaced by the EM-specific Superhuman backbone. (3) Architectures with a shared encoder, represented by DS2F [29].

Quantitative results. As shown in Tables 1, 2, and 3, our framework consistently outperforms existing methods across multiple datasets and diverse degradation settings, achieving superior performance on both SR and segmentation metrics. While jointly training IART and Superhuman improves segmentation accuracy at the cost of degraded SR quality, indicating a clear trade-off between the two tasks, our framework effectively mitigates this issue. Specifically, it surpasses single-task SR models in reconstruction fidelity while simultaneously outperforming existing joint frameworks in segmentation performance. Split/merge decompositions of VOI and repeated-inference results are provided in Appendices C and D, respectively.

Qualitative results. As illustrated in Figure 5, under the $8\times$ super-resolution setting, our framework achieves noticeably higher reconstruction fidelity and segmentation accuracy compared with other methods. When facing the more challenging $16\times$ super-resolution scenario, as shown in Figure 6, our method preserves cellular structures with the highest fidelity. While super-resolution at such a large scale inevitably introduces reconstruction inaccuracies for all methods, our framework still yields substantially more accurate segmentation results than competing methods. Additional comparisons and 3D segmentation visualizations are provided in Appendix H.

Comparison of different joint frameworks. We further compare our framework with representative joint frameworks in terms of segmentation performance, including a shared-encoder framework (DS2F [29]) and a cascaded framework (SR4IR [15]), as illustrated in Figure 7. For the shared-encoder framework, due to the structural mismatch between SR and EM segmentation networks as well as the difficulty of accurately upsampling affinity-based predictions, the predicted affinity maps tend to be blurred, with coarse and discontinuous boundaries, leading to degraded segmentation results. The cascaded framework produces more reasonable affinity predictions overall, but still fails on challenging structures, such as thin and elongated neuron instances shown in the figure. In

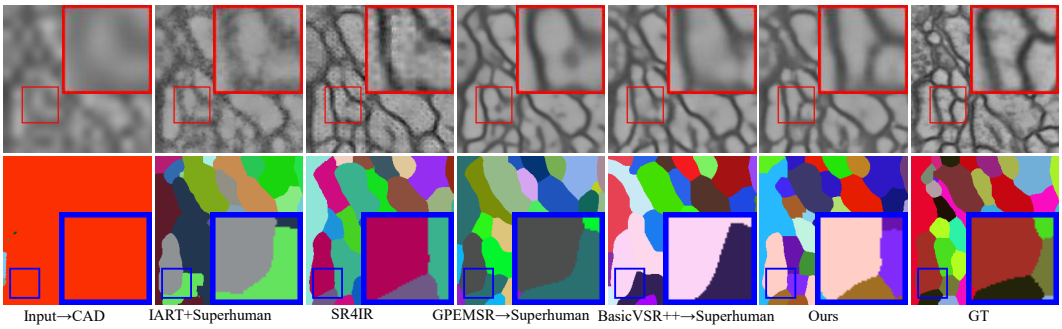

Figure 6: Qualitative comparison of $16\times$ bicubic-downsampled super-resolution and segmentation results on the CREMI C dataset. "+" indicates joint optimization of SR and segmentation, whereas "$\rightarrow$" denotes a cascaded framework with separate training. The input is obtained by bicubic interpolation of the LR image. We select methods with the best quantitative performance for comparison.

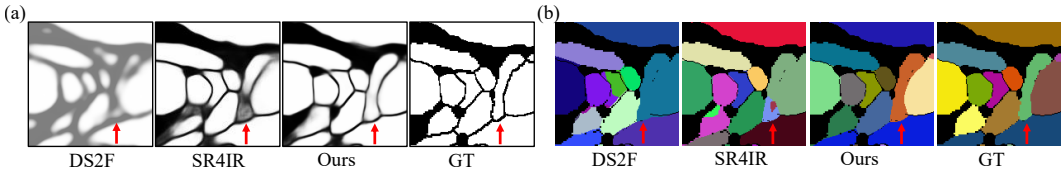

Figure 7: Qualitative comparison of segmentation performance across different joint frameworks. (a) Horizontal affinity maps. (b) The corresponding segmentation results.

contrast, our framework yields more accurate affinity maps and segmentation results, demonstrating its effectiveness in handling fine-scale and structurally complex cellular regions.

4.3 Ablation Studies

The effectiveness of the BWA module. We compare the proposed BWA module with the widely adopted second-order alignment-and-propagation strategy for inter-section feature alignment. Specifically, for the pairwise section alignment within the second-order framework, we adopt the state-of-the-art implicit alignment (IA) module [42] as the baseline. As shown in Rows 1–2 of Table 4, our BWA module achieves superior super-resolution performance. We further verify the BWA configuration in Table 5: enlarging the fusion range raises computational cost without measurable SR gain, and modifying the window overlap degrades SR, confirming that fusing only with the nearest neighboring sections is optimal.

The effectiveness of the semantic prior. By comparing Rows 2 and 3 as well as Rows 4 and 5 in Table 4, we observe that, for both CNN- and Transformer-based backbones, introducing an additional semantic prior into the SR network consistently improves SR performance. This trend is further corroborated by the qualitative results in Figure 8 (a), where the BWA module enhances reconstruction fidelity, and the incorporation of a semantic prior yields additional gains. In particular, the semantic prior enables more accurate reconstruction of fine cellular boundaries, especially for low-resolution inputs with fine-scale cellular structures.

The effectiveness of the structural prior. As shown in Table 6 and Figure 8 (b), introducing a structural prior learned from the partial diffusion model consistently improves segmentation performance on super-resolved images. Compared with an MLP that directly regresses P_{HR} from the SR embedding, our partial-diffusion prior also achieves lower VOI and ARAND on both volumes. Real, which feeds the ground-truth P_{HR} instead, is listed as a reference for the remaining gap.

Comparison of Partial and Full Diffusion. As shown in Table 6, Partial-4 provides comparable segmentation quality to Full-50 with $12.5\times$ fewer denoising steps. It achieves lower ARAND on both volumes and lower VOI on Wafer 36-2, while its Wafer 4 VOI is slightly higher. Thus, the main advantage is the reduction in sampling steps, with improvements in three of the four reported metrics.

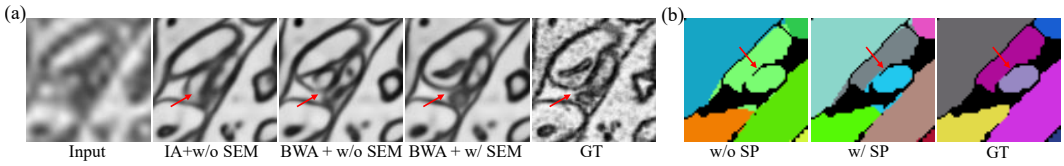

Figure 8: (a) Qualitative comparison illustrating the contributions of the BWA module and semantic prior to SR performance. (b) Qualitative comparison demonstrating the effectiveness of the structural prior on segmentation results. SEM denotes semantic prior, and SP denotes structural prior.

Table 4: Ablation on the effectiveness of the BWA module and the semantic prior. Trans denotes transformer backbone and SEM indicates whether the semantic prior is applied. The best result for each backbone is highlighted in **bold**.

Backbone	Alignment	SEM	Wafer 4		Wafer 36-2	
			PSNR↑	SSIM↑	PSNR↑	SSIM↑
CNN	IA	✗	17.50	0.4162	17.57	0.4831
CNN	BWA	✗	17.81	0.4421	18.04	0.5234
CNN	BWA	✓	**17.99**	**0.4466**	**18.28**	**0.5327**
Trans	BWA	✗	17.91	0.4416	18.14	0.5208
Trans	BWA	✓	**18.06**	**0.4557**	**18.38**	**0.5420**

Table 5: Ablation on BWA sliding window. Size denotes the fusion range, and Stride controls the overlap among windows. Our choice.

Size	Stride	Wafer 4		Wafer 36-2	
		PSNR↑	SSIM↑	PSNR↑	SSIM↑
3	3	17.92	0.4461	18.20	0.5305
3	1	**17.99**	0.4466	**18.28**	0.5327
5	1	17.96	**0.4469**	18.24	**0.5332**

Table 6: Ablation on the effectiveness of the structural prior. MLP directly regresses the HR prior from the SR embedding. Full and Partial denote the diffusion model with full and partial noise injection, respectively. Truncated indicates that the structural prior is obtained via truncated diffusion, while Real uses the real HR prior, which is inaccessible in practice and serves as an upper bound on performance. The best result is highlighted in **bold**.

Methods	Steps	Wafer 4		Wafer 36-2	
		VOI↓	ARAND↓	VOI↓	ARAND↓
w/o diff	-	1.5770	0.4431	0.6546	0.1038
MLP	-	1.5490	0.3785	0.6177	0.1056
Full	50	1.3623	0.1751	0.6013	0.0951
Truncated	4	1.4684	0.1496	0.6175	**0.0933**
Truncated	10	1.3794	0.1452	0.6369	0.1316
Partial	1	1.4644	0.2376	0.6347	0.1140
Partial	2	1.4409	0.1610	0.6168	0.1015
Partial	4	1.3726	**0.1230**	**0.5971**	0.0934
Partial	8	**1.3496**	0.1657	0.6198	0.1097
Real	-	1.2762	0.1275	0.5766	0.0896

Among the partial schedules, eight steps further reduce Wafer 4 VOI but worsen its ARAND and both metrics on Wafer 36-2. We select four steps to balance sampling cost and segmentation quality.

Comparison with truncated diffusion. A natural alternative for obtaining the structural prior is the truncated diffusion probabilistic model (TDPM) [45], which truncates the diffusion chain and uses a GAN to sample the chain's starting point in the prior space. For a fair comparison, we instantiate both the GAN and the diffusion backbone with the linear-layer architecture. Under the same starting timestep and noise schedule, we train two TDPM variants and report the comparison in Table 6: TDPM attains only a negligible margin on a single metric, while our partial-diffusion prior outperforms it on all remaining metrics without requiring TDPM's additional GAN component.

5 Conclusion

In this paper, we presented a unified framework for jointly performing EM image super-resolution and segmentation. On the super-resolution side, we leverage an auxiliary segmentation branch to inject a semantic prior that enhances reconstruction fidelity, and further introduce a BWA module to effectively exploit inter-section correlations. On the segmentation side, our key insight is that a lightweight partial diffusion model initialized from a noised SR embedding rather than Gaussian noise produces a structural prior that closely approximates real HR features and improves segmentation accuracy. Extensive experiments on multiple EM datasets under diverse degradation settings demonstrate that our framework simultaneously improves reconstruction fidelity and segmentation accuracy, supporting scalable connectomics analysis from rapid EM acquisition.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant 624B2137. The authors declare no competing interests.

References

[1] L. F. Abbott, D. D. Bock, E. M. Callaway, W. Denk, C. Dulac, A. L. Fairhall, I. Fiete, K. M. Harris, M. Helmstaedter, V. Jain, et al. The mind of a mouse. *Cell*, 182(6):1372–1376, 2020.

[2] I. Arganda-Carreras, S. C. Turaga, D. R. Berger, D. Cireşan, A. Giusti, L. M. Gambardella, J. Schmidhuber, D. Laptev, S. Dwivedi, J. M. Buhmann, et al. Crowdsourcing the creation of image segmentation algorithms for connectomics. *Frontiers in neuroanatomy*, 9:152591, 2015.

[3] T. Beier, C. Pape, N. Rahaman, T. Prange, S. Berg, D. D. Bock, A. Cardona, G. W. Knott, S. M. Plaza, L. K. Scheffer, et al. Multicut brings automated neurite segmentation closer to human performance. *Nature methods*, 14(2):101–102, 2017.

[4] K. C. Chan, S. Zhou, X. Xu, and C. C. Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5972–5981, 2022.

[5] I. Chen, W.-T. Chen, Y.-W. Liu, Y.-C. Chiang, S.-Y. Kuo, M.-H. Yang, et al. Unirestore: Unified perceptual and task-oriented image restoration model using diffusion prior. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17969–17979, 2025.

[6] Y. Chen, H. Shi, X. Liu, T. Shi, R. Zhang, D. Liu, Z. Xiong, and F. Wu. Tokenunify: Scaling up autoregressive pretraining for neuron segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13604–13613, 2025.

[7] K. De Haan, Z. S. Ballard, Y. Rivenson, Y. Wu, and A. Ozcan. Resolution enhancement in scanning electron microscopy using deep learning. *Scientific reports*, 9(1):12050, 2019.

[8] S. Dorkenwald, C. M. Schneider-Mizell, D. Brittain, A. Halageri, C. Jordan, N. Kemnitz, M. A. Castro, W. Silversmith, J. Maitin-Shephard, J. Troidl, et al. Cave: Connectome annotation versioning engine. *Nature methods*, 22(5):1112–1120, 2025.

[9] L. Fang, F. Monroe, S. W. Novak, L. Kirk, C. R. Schiavon, S. B. Yu, T. Zhang, M. Wu, K. Kastner, A. A. Latif, et al. Deep learning-based point-scanning super-resolution imaging. *Nature Methods*, 18(4):406–416, 2021.

[10] J. Funke, S. Saalfeld, D. Bock, S. Turaga, and E. Perlman. Miccai challenge on circuit reconstruction from electron microscopy images, 2016.

[11] J. Funke, F. Tschopp, W. Grisaitis, A. Sheridan, C. Singh, S. Saalfeld, and S. C. Turaga. Large scale image segmentation with structured loss based deep learning for connectome reconstruction. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1669–1680, 2018.

[12] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *The journal of machine learning research*, 13(1):723–773, 2012.

[13] M. Haris, G. Shakhnarovich, and N. Ukita. Task-driven super resolution: Object detection in low-resolution images. In *International conference on neural information processing*, pages 387–395. Springer, 2021.

[14] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[15] J. Kim, J. Oh, and K. M. Lee. Beyond image super-resolution for image recognition with task-driven perceptual loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2651–2661, 2024.

[16] J. Kim, J. Oh, and K. M. Lee. Exploiting diffusion prior for task-driven image restoration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10151–10161, 2025.

[17] D. P. Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[18] K. Lee, J. Zung, P. Li, V. Jain, and H. S. Seung. Superhuman accuracy on the snemi3d connectomics challenge. *arXiv preprint arXiv:1706.00120*, 2017.

[19] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021.

[20] J. Liang, Y. Fan, X. Xiang, R. Ranjan, E. Ilg, S. Green, J. Cao, K. Zhang, R. Timofte, and L. V. Gool. Recurrent video restoration transformer with guided deformable attention. *Advances in Neural Information Processing Systems*, 35:378–393, 2022.

[21] Q. Liu, C. Zhuang, P. Gao, and J. Qin. Cdformer: When degradation prediction embraces diffusion model for blind image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7455–7464, 2024.

[22] X. Liu, M. Cai, Y. Chen, Y. Zhang, T. Shi, R. Zhang, X. Chen, and Z. Xiong. Cross-dimension affinity distillation for 3d em neuron segmentation. In *2024 IEEE/CVF Conference on Computer*

Vision and Pattern Recognition (CVPR), pages 11104–11113. IEEE Computer Society, 2024.
[23] I. Loshchilov and F. Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.
[24] C. Lu, K. Chen, H. Qiu, X. Chen, G. Chen, X. Qi, and H. Jiang. Diffusion-based deep learning method for augmenting ultrastructural imaging and volume electron microscopy. *Nature Communications*, 15(1):4677, 2024.
[25] A. Q. Nichol and P. Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021.
[26] J. Nunez-Iglesias, R. Kennedy, T. Parag, J. Shi, and D. B. Chklovskii. Machine learning of hierarchical clustering to segment 2d and 3d images. *PloS one*, 8(8):e71715, 2013.
[27] S. M. Plaza, L. K. Scheffer, and D. B. Chklovskii. Toward large-scale connectome reconstructions. *Current opinion in neurobiology*, 25:201–210, 2014.
[28] Y. Qian, J. Xu, L. F. Drummy, and Y. Ding. Effective super-resolution methods for paired electron microscopic images. *IEEE Transactions on Image Processing*, 29:7317–7330, 2020.
[29] Z. Qiu, Y. Hu, X. Chen, D. Zeng, Q. Hu, and J. Liu. Rethinking dual-stream super-resolution semantic learning in medical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(1):451–464, 2023.
[30] M. S. Rad, B. Bozorgtabar, C. Musat, U.-V. Marti, M. Basler, H. K. Ekenel, and J.-P. Thiran. Benefiting from multitask learning to improve single image super-resolution. *Neurocomputing*, 398:304–313, 2020.
[31] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. 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.
[32] A. Shapson-Coe, M. Januszewski, D. R. Berger, A. Pope, Y. Wu, T. Blakely, R. L. Schalek, P. H. Li, S. Wang, J. Maitin-Shepard, et al. A petavoxel fragment of human cerebral cortex reconstructed at nanoscale resolution. *Science*, 384(6696):eadk4858, 2024.
[33] S. Shi, J. Gu, L. Xie, X. Wang, Y. Yang, and C. Dong. Rethinking alignment in video super-resolution transformers. *Advances in Neural Information Processing Systems*, 35:36081–36093, 2022.
[34] J. Shou, Z. Xiao, S. Deng, W. Huang, P. Shi, R. Zhang, Z. Xiong, and F. Wu. Learning large-factor em image super-resolution with generative priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11313–11322, 2024.
[35] H. Wang, L. Lin, H. Hu, Q. Chen, Y. Li, Y. Iwamoto, X.-H. Han, Y.-W. Chen, and R. Tong. Patch-free 3d medical image segmentation driven by super-resolution technique and self-supervised guidance. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 131–141. Springer, 2021.
[36] L. Wang, D. Li, Y. Zhu, L. Tian, and Y. Shan. Dual super-resolution learning for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3774–3783, 2020.
[37] Y. Wei, Z. Zhang, H. Zheng, R. Hong, Y. Yang, and M. Wang. Sginet: Toward sufficient interaction between single image deraining and semantic segmentation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6202–6210, 2022.
[38] R. Wu, L. Sun, Z. Ma, and L. Zhang. One-step effective diffusion network for real-world image super-resolution. *Advances in Neural Information Processing Systems*, 37:92529–92553, 2024.
[39] R. Wu, T. Yang, L. Sun, Z. Zhang, S. Li, and L. Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 25456–25467, 2024.
[40] Y. Wu, C. Pan, G. Wang, Y. Yang, J. Wei, C. Li, and H. T. Shen. Learning semantic-aware knowledge guidance for low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1662–1671, 2023.
[41] Y. Xie, Y. Ding, and S. Ji. Augmented equivariant attention networks for microscopy image transformation. *IEEE Transactions on Medical Imaging*, 41(11):3194–3206, 2022.
[42] K. Xu, Z. Yu, X. Wang, M. B. Mi, and A. Yao. Enhancing video super-resolution via implicit resampling-based alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2546–2555, 2024.
[43] Q. Zhang, X. Liu, W. Li, H. Chen, J. Liu, J. Hu, Z. Xiong, C. Yuan, and Y. Wang. Distilling semantic priors from sam to efficient image restoration models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25409–25419, 2024.
[44] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 286–301, 2018.

[45] H. Zheng, P. He, W. Chen, and M. Zhou. Truncated diffusion probabilistic models and diffusion-based adversarial auto-encoders. *arXiv preprint arXiv:2202.09671*, 2022.

[46] X. Zhu, H. Hu, S. Lin, and J. Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9308–9316, 2019.

[47] Y. Zhu, R. Wang, S. Lu, H. Yan, J. Li, and K. Zhang. One-step flow for image super-resolution with tunable fidelity-realism trade-offs. In *International Conference on Learning Representations*, volume 2026, pages 91476–91505, 2026.

Appendix

A Backbone Implementation Details

A.1 Auxiliary Network for the Semantic Prior

The auxiliary network used to produce the semantic prior comprises three modules: a shared encoder, an SR decoder, and a segmentation decoder. All three adopt the architecture of Superhuman [18], the anisotropic 3D U-Net widely used for EM neuron segmentation. Specifically, the shared encoder reuses Superhuman's encoder, while both the SR decoder and the segmentation decoder reuse Superhuman's decoder architecture.

A.2 SR Encoder and HR Encoder

The SR encoder and HR encoder in the segmentation part both take a single EM section as input and output a d-dimensional prior vector. The two encoders share the same architecture but maintain separate parameters: the HR encoder produces P_{HR} from a real HR EM section in Stage I (and is frozen in Stage II), while the SR encoder produces P_{SR} from a super-resolved EM section in Stage II as the input to the partial diffusion process.

The encoder consists of three stages. The input EM section is first downsampled via a single convolution. A stack of residual blocks then refines the features at constant resolution, after which two convolutions progressively expand the channel dimension to d. Finally, global average pooling collapses the spatial dimensions, and an MLP produces the final d-dimensional prior vector.

A.3 Prior Fusion Module

The Prior Fusion module is used in both Stage I (for HR-prior injection) and Stage II (for structural-prior injection). It is implemented as a feature-wise affine modulation: a single linear layer (without bias) projects the d-dimensional prior vector P into channel-wise scale and shift parameters that modulate the segmentation feature $F_{\mathrm{seg}} \in \mathbb{R}^{C\times H\times W}$ for each EM section:

$$[\gamma, \beta] = W_P\, P, \qquad F_{\mathrm{out}} = \gamma \otimes F_{\mathrm{seg}} + \beta, \tag{8}$$

where $W_P \in \mathbb{R}^{2C\times d}$ is the learnable projection, $\gamma, \beta \in \mathbb{R}^C$ are obtained by splitting the projected vector along the channel dimension and broadcast over spatial locations, and $\otimes$ denotes element-wise multiplication. The prior P corresponds to P_{HR} in Stage I and $\hat{P}_{\mathrm{SR}}$ in Stage II.

B Computational Complexity Analysis

Quantitative comparisons of computational complexity are presented in Tables 7 and 8. Our framework exhibits lower computational complexity than most existing methods.

C Background on EM Connectomics and Segmentation Errors

EM connectomics reconstructs neural circuits by tracing individual neurons through serial EM sections and identifying their synaptic connections. Neuron instance segmentation assigns a consistent identity to each neuron throughout the volume, providing the basis for assigning synapses to the correct neurons and recovering circuit connectivity. Preserving neuronal continuity and separating adjacent neurons are therefore essential for connectome reconstruction.

VOI is the sum of a split term $\mathrm{VOI}_{\mathrm{s}}$, which penalizes false splits (over-segmentation), and a merge term $\mathrm{VOI}_{\mathrm{m}}$, which penalizes false merges (under-segmentation). The two errors are not equivalent for connectomics, where the segmentation defines the wiring diagram and every neuron must be recovered as exactly one segment. A split fragments a neuron and under-reports its connections, and is fixed by reconnecting the fragments. A merge fuses distinct neurons, creates connections that do not exist, and may require several split operations to undo when two neurons touch at multiple places; since automated segmentations are corrected by human proofreaders, which can take years on

Table 7: Computational complexity comparison for joint $8\times$ super-resolution and segmentation. Ours(CNN) denotes the model with a CNN backbone evaluated on the MEC dataset, while Ours(Trans) denotes the model with a Transformer backbone evaluated on the CREMI A and B datasets. FLOPs are computed with an input of 9 sections, where the ground-truth resolution is 1024×1024. Inference time is measured on an NVIDIA A100. "+" indicates that the super-resolution and segmentation networks are jointly optimized. For methods in which the SR and segmentation models are trained separately, segmentation inference is performed by partitioning the image into patches and stitching the predicted affinities, following the common practice of affinity-based segmentation methods. Other methods do not involve affinity stitching during segmentation inference.

Methods		Params/M	FLOPs/G	Time/ms
Super-resolution	Segmentation			
BasicVSR++	Superhuman	9.02	9512.06	1495
BasicVSR++	CAD	8.37	2296.98	357
GPEMSR	Superhuman	70.80	73335.29	4396
GPEMSR	CAD	70.14	66120.22	3259
IART	Superhuman	14.06	13502.57	4197
IART	CAD	13.40	6287.50	3060
OFTSR	Superhuman	95.11	51407.90	6623
OFTSR	CAD	94.46	44192.82	5420
IART + Superhuman		14.02	7831.67	3181
DS2F		20.81	179.30	21
SR4IR		13.52	4392.67	1011
Ours(CNN)		8.65	4204.24	609
Ours(Trans)		11.14	5821.92	2772

Table 8: Computational complexity comparison for joint $16\times$ super-resolution and segmentation. Ours adopts a Transformer backbone. FLOPs are computed with an input of 9 sections, where the ground-truth resolution is 1024×1024. Inference time is measured on an NVIDIA A100. "+" indicates that the super-resolution and segmentation networks are jointly optimized. For methods in which the SR and segmentation models are trained separately, segmentation inference is performed by partitioning the image into patches and stitching the predicted affinities, following the common practice of affinity-based segmentation methods. Other methods do not involve affinity stitching during segmentation inference.

Methods		Params/M	FLOPs/G	Time/ms
Super-resolution	Segmentation			
BasicVSR++	Superhuman	9.47	8825.36	1482
BasicVSR++	CAD	8.81	1610.29	347
GPEMSR	Superhuman	74.32	65986.22	4112
GPEMSR	CAD	73.67	58771.15	2977
IART	Superhuman	14.20	9909.96	2191
IART	CAD	13.55	2694.89	1057
IART + Superhuman		14.17	4239.05	1180
DS2F		20.95	80.16	18
SR4IR		13.67	3118.31	722
Ours		11.29	4352.79	1197

a large dataset [8], merges are the costlier error. We report both terms for our framework and two representative joint frameworks, IART + Superhuman and SR4IR, in Table 9.

Our improvement in VOI is accompanied by consistent reductions in merge errors, the more costly error type in connectome reconstruction. Our framework achieves the lowest merge term and total VOI among the compared methods across all seven evaluated settings.

Table 9: Split/merge decomposition of VOI under the settings of Tables 1, 2, and 3. VOI_s and VOI_m denote the split and merge terms, respectively. The $4\times$ setting includes beam blur and shot noise. All methods use a fixed agglomeration threshold of 0.5. The best and second-best results among the compared methods are highlighted in **bold** and underline, respectively. Components and totals are rounded independently.

Volume	IART + Superhuman			SR4IR			Ours		
	$\mathrm{VOI}_s \downarrow$	$\mathrm{VOI}_m \downarrow$	VOI$\downarrow$	$\mathrm{VOI}_s \downarrow$	$\mathrm{VOI}_m \downarrow$	VOI$\downarrow$	$\mathrm{VOI}_s \downarrow$	$\mathrm{VOI}_m \downarrow$	VOI$\downarrow$
Wafer 4 ($8\times$)	1.0596	0.4760	1.5356	1.2103	0.3962	1.6064	**0.9843**	**0.3883**	**1.3726**
Wafer 36-2 ($8\times$)	0.5030	0.4726	0.9756	0.4975	0.2726	0.7701	**0.3708**	**0.2263**	**0.5971**
CREMI A ($8\times$)	0.3919	0.3233	0.7152	0.3830	0.3439	0.7270	**0.3739**	**0.2978**	**0.6717**
CREMI B ($8\times$)	**0.7199**	1.1165	1.8364	0.8567	0.8494	1.7061	0.8412	**0.7920**	**1.6332**
CREMI C ($16\times$)	1.4776	1.5603	3.0379	**1.3647**	1.4232	2.7879	1.5071	**0.9810**	**2.4881**
Wafer 4 ($4\times$)	**1.3016**	0.5754	1.8770	1.5087	0.3272	1.8360	1.3715	**0.2566**	**1.6280**
Wafer 36-2 ($4\times$)	0.3886	0.3748	0.7633	0.3635	0.2589	0.6224	**0.3400**	**0.2171**	**0.5572**

D Run-to-Run Variability

The only stochastic component of our pipeline at inference is the partial diffusion process that generates the structural prior, i.e., the noise added to the SR embedding at the starting step and the noise injected at each denoising step. The SR branch is deterministic, so PSNR and SSIM are identical across runs and only the segmentation metrics vary. We repeat inference five times using the same trained checkpoint and report the mean and standard deviation of VOI and ARAND in Table 10. The main-paper tables retain the single-run results.

Table 10: Run-to-run variability over five inference runs under the settings of Tables 1, 2, and 3. We report the mean and standard deviation of VOI and ARAND. The $4\times$ setting includes beam blur and shot noise. Best baseline denotes the lowest baseline VOI in the corresponding main-paper table. Δ is the baseline VOI minus our mean VOI, and σ is our VOI standard deviation. The last column expresses this margin in units of σ.

Volume	Ours (mean $\pm$ std)		VOI comparison		
	VOI$\downarrow$	ARAND$\downarrow$	Best baseline	Δ	Δ/σ
Wafer 4 ($8\times$)	1.3790 ± 0.0309	0.1398 ± 0.0251	1.5356	0.1566	$5.1\times$
Wafer 36-2 ($8\times$)	0.5685 ± 0.0179	0.0806 ± 0.0093	0.7701	0.2016	$11.3\times$
CREMI A ($8\times$)	0.6739 ± 0.0065	0.0895 ± 0.0025	0.7152	0.0413	$6.4\times$
CREMI B ($8\times$)	1.6443 ± 0.0263	0.1136 ± 0.0083	1.7061	0.0618	$2.3\times$
CREMI C ($16\times$)	2.4846 ± 0.0134	0.2107 ± 0.0026	2.7879	0.3033	$22.6\times$
Wafer 4 ($4\times$)	1.6256 ± 0.0190	0.3269 ± 0.0120	1.8360	0.2104	$11.1\times$
Wafer 36-2 ($4\times$)	0.5287 ± 0.0190	0.1010 ± 0.0100	0.6224	0.0937	$4.9\times$

Our mean VOI is lower than the best baseline VOI across all seven evaluated settings, with margins ranging from 2.3 to 22.6 times the inference standard deviation. In every setting, our mean VOI plus twice its standard deviation remains below the best baseline VOI. These comparisons quantify the performance margin relative to the sampling variability of a fixed checkpoint.

E SR Backbones: Configurations and Full Comparison

The CNN variant adopts the backbone of GPEMSR [34] and fuses the aligned features in the BWA module with its 3DA module, while the Transformer variant adopts the backbone of IART [42] and performs the fusion with a 3D Swin Transformer [33]. Since the two variants share their backbones with GPEMSR and IART, a gain over either baseline cannot be attributed to a stronger backbone or a larger capacity. Table 11 reports both variants on all five test volumes under the settings of Tables 1 and 2.

Two observations follow. (i) Neither backbone dominates. The Transformer variant reconstructs better on MEC and CREMI A, whereas the CNN variant gives lower VOI and ARAND on all five volumes and also reconstructs better under noise and $16\times$ downsampling. (ii) Both variants achieve lower VOI and ARAND than every baseline reported in Tables 1 and 2 across all five volumes.

Table 11: Both backbones of our framework on the five volumes of Tables 1 and 2. The better of the two variants is in **bold**; † marks the variant reported in the main paper.

Dataset	Ours (CNN)				Ours (Transformer)			
	PSNR↑	SSIM↑	VOI↓	ARAND↓	PSNR↑	SSIM↑	VOI↓	ARAND↓
Wafer 4	17.99†	0.4466†	**1.3726**†	**0.1230**†	**18.06**	**0.4557**	1.4022	0.1372
Wafer 36-2	18.28†	0.5327†	**0.5971**†	**0.0934**†	**18.38**	**0.5420**	0.6081	0.1014
CREMI A	31.00	0.8286	**0.6689**	**0.0809**	**31.13**†	**0.8318**†	0.6717†	0.0893†
CREMI B	**25.18**	**0.6741**	**1.4822**	**0.0960**	25.13†	0.6724†	1.6332†	0.1038†
CREMI C	**23.08**	**0.5738**	**2.3188**	**0.1982**	23.01†	0.5645†	2.4881†	0.2083†

F Limitations

While our framework consistently outperforms existing methods on both super-resolution and segmentation across multiple EM datasets, several limitations remain. First, the framework relies on a two-stage training scheme by design: the HR prior must be shaped in Stage I before the diffusion model in Stage II can be trained to approximate it. Exploring training schedules that further reduce overall training cost is a possible direction for future work. Second, the HR prior is obtained from real HR EM images during training; relaxing this requirement through semi-supervised or self-supervised priors is a natural extension for settings where paired HR data are unavailable. Third, reconstructing fine structural details and accuratelysegmenting neurons remain challenging under severe downsampling. As illustrated in Figure 9, the predicted segmentation splits a single ground-truth neuron into multiple regions.

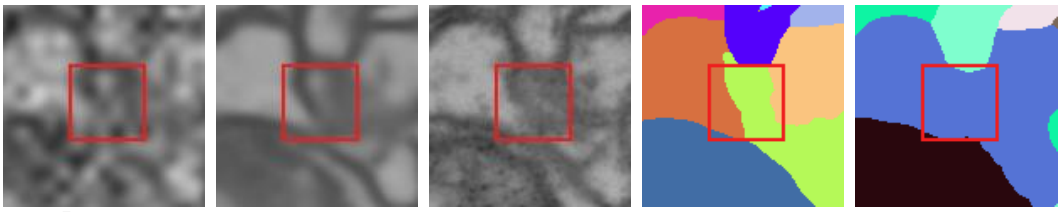

Figure 9: A failure case on CREMI B under the noisy $8\times$ super-resolution setting.

G Broader Impact

Our framework targets the connectomics pipeline, where reducing EM acquisition time and producing accurate neuron segmentations together enable larger-scale brain mapping and faster hypothesis testing in neuroscience research. By improving reconstruction fidelity from low-resolution scans while preserving downstream segmentation accuracy, the framework can lower the imaging time and storage cost required for cubic-millimeter or larger volume reconstructions, contributing to the long-term goal of large-scale connectomics.

We do not foresee direct negative societal impacts of this work. As with any deep learning system, the framework may produce reconstruction or segmentation errors that can propagate to downstream analyses; results should therefore be cross-validated by domain experts before being used to draw biological conclusions.

H More Visual Comparisons

More qualitative comparison of super-resolution and segmentation. As shown in Figure 10, our framework consistently achieves superior performance in both super-resolution and segmentation under various degradation settings, compared with existing methods.

3D segmentation visualizations. As shown in Figure 11, our framework is able to better preserve the spatial integrity of neuronal structures under various degradation conditions, which is beneficial for downstream neuron connectivity analysis.

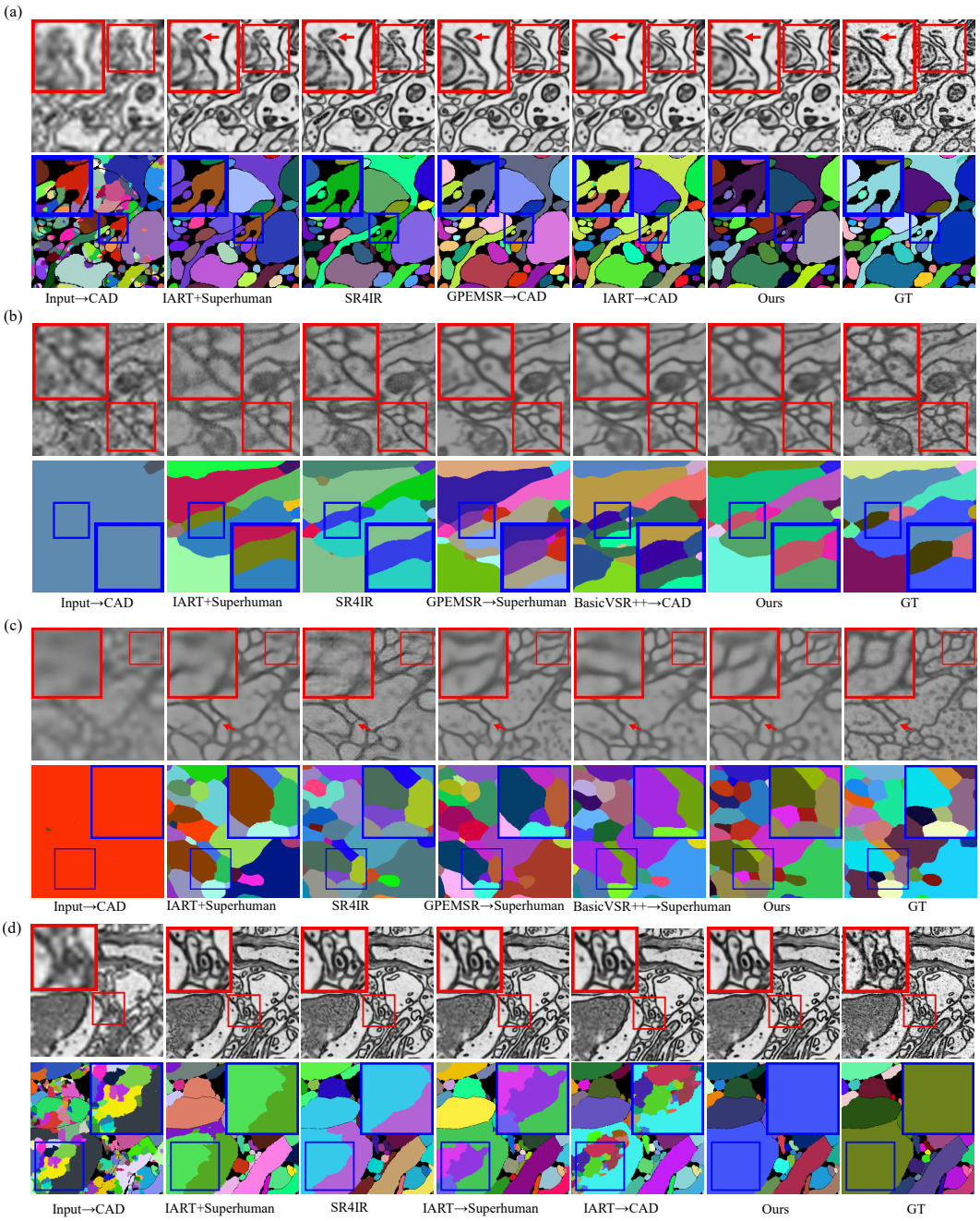

Figure 10: Qualitative comparison of super-resolution and segmentation results. "+" denotes joint optimization of super-resolution and segmentation, whereas "→" indicates a cascaded framework with separately trained SR and segmentation models. The input is obtained by bicubic interpolation of the low-resolution image. For visual comparison, we select methods with the best quantitative performance. (a) $8\times$ super-resolution on the wafer 36-2 dataset. (b) SR ($8\times$) with additive Gaussian noise ($\sigma \in [10, 25]$) on the CREMI B dataset. (c) $16\times$ super-resolution on the CREMI C dataset. (d) $4\times$ super-resolution with beam blur and shot noise on the wafer 4 dataset.

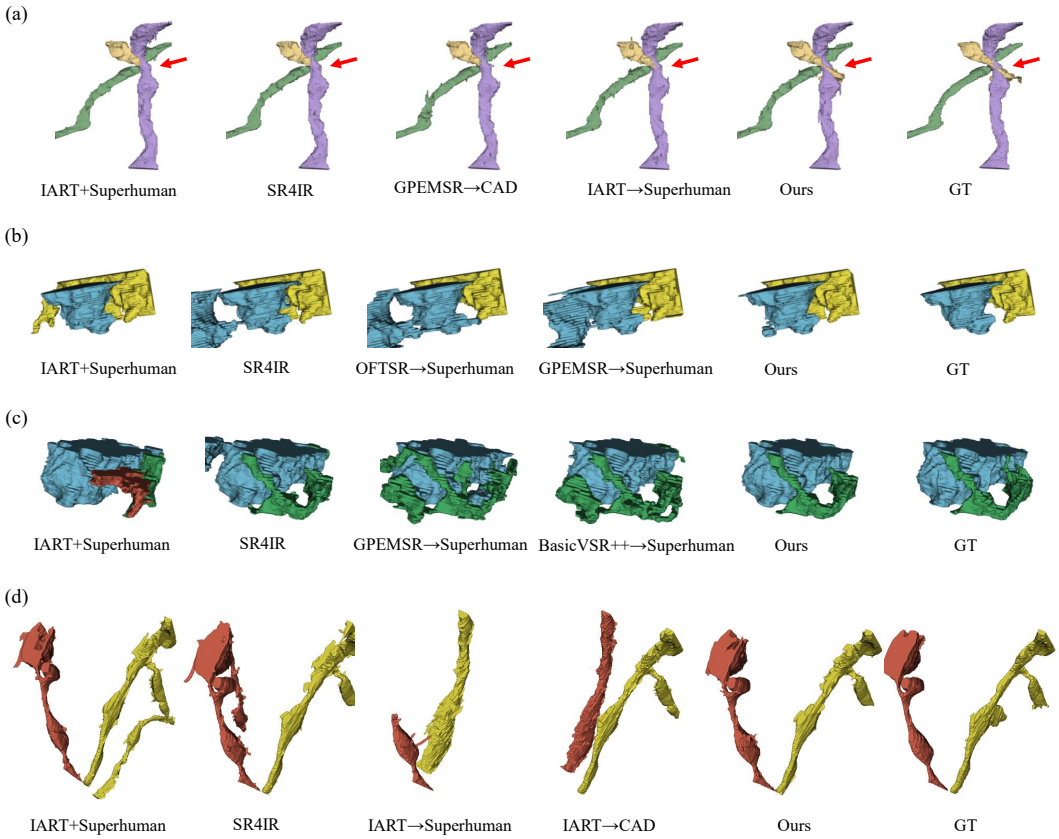

Figure 11: Qualitative comparison of 3D segmentation visualizations. "+" denotes joint optimization of super-resolution and segmentation, whereas "→" indicates a cascaded framework with separately trained SR and segmentation models. For visual comparison, we select methods with the best quantitative performance. (a) $8\times$ super-resolution on the wafer 4 dataset. (b) SR $(8\times)$ with additive Gaussian noise ($\sigma \in [10, 25]$) on the CREMI B dataset. (c) $16\times$ super-resolution on the CREMI C dataset. (d) $4\times$ super-resolution with beam blur and shot noise on the wafer 4 dataset.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We clearly summarize the contributions at the end of the introduction and briefly describe them in the abstract.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Appendix F.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide sufficient details to reproduce the paper, including the architecture of each module and hyperparameters selection. The code is publicly available.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code is publicly available.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: See Section 4.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: See Appendix D.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See Section 4.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in this paper fully complies with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: See Appendix G.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All external assets used in our paper are properly cited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We will introduce the new code as part of this work. These assets will be publicly released after the paper is accepted. We will provide documentation on the use of the code and models along with the release to ensure reproducibility and transparency.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.