SkelEM: Training-Signal Decoupling of Skeleton and Diffusion for Self-supervised Axial Super-Resolution in Volume Microscopy

Bohao Chen[1,2], Yanchao Zhang[2,3], Yanan Lv[2,4], Chenxun Deng[2,4], Hua Han[2,3,4], and Xi Chen[2,✉]

[1] School of Advanced Interdisciplinary Sciences, University of Chinese Academy of Sciences, China
[2] State Key Laboratory of Brain Cognition and Brain-inspired Intelligence Technology, Institute of Automation, Chinese Academy of Sciences, China
{chenbohao2024, zhangyanchao2021, lvyanan2018, dengchenxun2025, hua.han, xi.chen}@ia.ac.cn
[3] School of Future Technology, University of Chinese Academy of Sciences, China
[4] School of Artificial Intelligence, University of Chinese Academy of Sciences, China

Abstract. Volume microscopy, including electron and light microscopy, suffers from severe anisotropic resolution due to physical axial sectioning. Existing self-supervised axial super-resolution (ASR) methods face a trilemma bounded by overly smoothed regression textures, structural hallucinations of pure diffusion models, and prohibitive inference latency. In this paper, we propose Skeleton-refinE Microscopy (SkelEM), a self-supervised framework that decouples ASR at the training-signal level: a frozen topological network and a diffusion refiner are optimized by disjoint objectives, separating low-frequency topology formulation from high-frequency detail enhancement. Building on this deterministic skeleton, we exploit a unified cycle-consistent mechanism on input sparse slices to simultaneously extract a real-domain residual prior and bidirectionally align the diffusion refiner, washing away cross-plane artifacts without synthetic bias. By truncating the reverse diffusion process with this physical prior, SkelEM achieves high-fidelity detail restoration in merely ≤ 5 steps. To rigorously assess cross-instrument generalization, we further introduce BRAVE-ASR, a new benchmark of co-aligned anisotropic and isotropic volumes acquired on a Plasma-FIB instrument. Across public benchmarks, SkelEM achieves the most favorable balance across the fidelity-perception trade-off among self-supervised methods, with state-of-the-art downstream membrane segmentation performance and robust zero-shot generalization across distinct modalities.

Keywords: Isotropic reconstruction · Diffusion models · Volume Microscopy

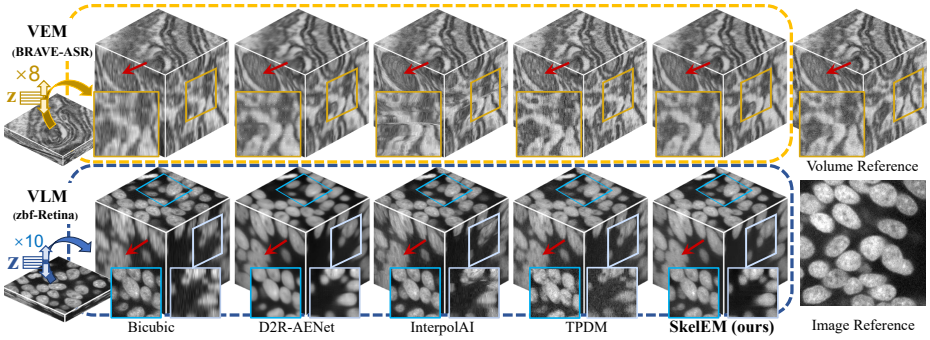

Fig. 1: SkelEM demonstrates robust axial super-resolution performance across different imaging modalities and instruments. **Top row (yellow box):** A challenging 8× axial super-resolution zero-shot instrument transfer task on our BRAVE-ASR dataset. **Bottom row (blue box):** A real-world (no high-resolution reference) 10× axial super-resolution task on a volume light microscopy dataset.

1 Introduction

Volume microscopy (VM) offers an unprecedented three-dimensional (3D) window into life's structure, from cellular organelles to complex tissue architectures [28,34]. However, its effectiveness across both electron and light modalities is critically hampered by anisotropic resolution [10, 44]. Most high-throughput VM techniques, such as serial block-face scanning electron microscopy (SBF-SEM) or confocal light microscopy, are limited by physical sectioning or optical physics to an axial resolution far lower than their lateral resolution. For instance, SBF-SEM achieves a lateral resolution of 4-10 nm but is limited to 25-40 nm axially [3], while standard confocal and two-photon microscopy are limited to 200-300 nm in lateral plane, versus 500-800 nm axially [8]. While methods like focused ion beam scanning electron microscopy (FIB-SEM) can achieve isotropic nanoscale resolution [43], their prohibitive cost and low throughput mean that the vast majority of large-volume datasets suffer from this axial deficiency [19]. The anisotropic resolution introduces structural and topological distortions [35] that can lead to erroneous circuit analysis in connectomics [5] and flawed morphological analysis in biology [26], directly impeding scientific discovery.

In recent years, computational approaches for axial super-resolution (ASR) have been developed to bridge this resolution gap, but face significant hurdles. Traditional interpolation produces blurry results, while supervised deep learning is impractical due to the extreme difficulty of acquiring isotropic 3D training volumes [10, 21]. Consequently, the field has shifted towards self-supervised methods. However, existing self-supervised paradigms face a fundamental fidelity-perception trade-off [2]. Regression-based 3D convolutional networks optimized via pixel-wise constraints inevitably suffer from regression-to-the-mean, producing overly smoothed volumes [4, 11], while generative methods such as diffusion models synthesize high-frequency details at the cost of structural hallucinations and prohibitive inference latency [22,23,39]. We attribute both failure modes to

a shared root cause: the absence of a faithful structural skeleton that captures the highly nonlinear deformations of biological ultrastructures. Without such guidance, regression models over-smooth and generative models hallucinate. A domain-adapted structural skeleton would not only anchor reconstruction topology, but also enable truncated diffusion [30, 45] to recover high-frequency details in very few steps, rendering the iterative denoising process both structurally reliable and computationally practical.

To this end, we propose Skeleton-refinE Microscopy (SkelEM), a self-supervised framework with two specialized stages decoupled at the training-signal level. SkelEM first establishes a deterministic structural skeleton via a topological network trained on domain-specific synthetic manifolds, deliberately discarding its detail refiner to enforce clean topology-texture separation. To recover missing biological textures without synthetic bias, a unified cycle-consistent mechanism on authentic sparse slices simultaneously extracts a real-domain residual prior and bidirectionally aligns a diffusion refiner to eliminate cross-plane artifacts. By initializing the reverse diffusion process from this physically grounded prior, SkelEM achieves high-fidelity reconstruction in merely ≤ 5 steps while strictly preserving 3D structural integrity.

In summary, our contributions are: (1) SkelEM, a two-stage self-supervised ASR framework with training-signal decoupling—the topological skeleton network and the diffusion refiner are optimized by disjoint objectives with no shared gradients—where a domain-adapted structural skeleton serves as the key enabling condition for effective truncated diffusion in the microscopy domain. (2) A unified cycle-consistent mechanism on authentic sparse slices that simultaneously eliminates synthetic bias and extracts a physically grounded residual prior, enabling high-fidelity diffusion refinement in ≤ 5 steps. (3) Extensive validation on FIB-SEM benchmarks and the newly introduced BRAVE-ASR dataset demonstrates that SkelEM achieves the most favorable balance across the fidelity-perception trilemma among self-supervised methods (see normalized rank score in Fig. 2), delivering best perceptual quality in XZ/YZ views alongside competitive fidelity and SOTA downstream segmentation performance. The framework further demonstrates applicability beyond volume electron microscopy (VEM), with qualitative validation on volume light microscopy (VLM) confirming its modality-agnostic design.

2 Related Work

Axial Super-Resolution from Anisotropic Volumes. Enhancing axial resolution in anisotropic 3D volumes is a fundamental challenge across VEM, VLM, and clinical imaging (e.g., CT/MRI). The extreme difficulty of acquiring paired isotropic ground-truth volumes has driven the field toward self-supervised methods [11], which leverage high-resolution lateral (XY) planes as an internal supervisory signal to enhance the low-resolution axial (Z) direction. Existing self-supervised ASR methods fall into three paradigms, each with a characteristic failure mode. (1) 2D super-resolution on orthogonal planes [6,16,21,32,33,39,40], pioneered by Weigert *et al.* [40] and extended with unsupervised degradation

learning [6], preserves lateral details but produces discontinuous cross-plane artifacts; (2) video frame interpolation-inspired methods [14, 18, 27, 42] ensure 3D continuity but struggle with the large nonlinear morphological changes of biological ultrastructures; (3) methods that refine 3D volumes by 2D diffusion models trained on XY planes [22,23,37] achieve strong perceptual quality but incur prohibitive computational costs [39]. Navigating this trilemma between perceptual quality [2], 3D structural consistency, and computational tractability remains an open challenge, motivating the training-signal decoupling strategy of our work.

Video Frame Interpolation. Video frame interpolation (VFI) generates intermediate frames between existing input frames and serves as a strong methodological analogue for ASR. A dominant paradigm, exemplified by Super-SloMo [17] and RIFE [13], explicitly estimates optical flow to warp and synthesize intermediate frames and continues to be refined for challenging non-linear motions [25]. Complementing these flow-based approaches, recent flow-free methods leverage state-space models such as Mamba [15] or diffusion-based models [9, 14]. While we adopt the flow estimation backbone of RIFE [13] as our topological skeleton network, a critical distinction from standard VFI pipelines is that we deliberately discard its detail refiner after pre-training. This design choice prevents the leakage of synthetic hallucinations inherent in pseudo-HR volumes into the texture generation stage, enforcing a strict separation between low-frequency topology and high-frequency detail recovery. Furthermore, as we demonstrate in Sec. 4.4, VFI models pre-trained on natural videos fail to capture the complex nonlinear deformations of biological ultrastructures, necessitating domain-specific retraining on synthetic microscopy manifolds.

Diffusion Models for Image Restoration. Diffusion probabilistic models (DPMs) have achieved state-of-the-art results in high-fidelity image restoration [7, 12], but their iterative denoising process renders them impractical for large-scale biological volumes [39]. While truncated diffusion strategies [36, 45] can accelerate inference by initializing from an informative prior, their effectiveness critically depends on the quality of that initialization. In the absence of a structurally reliable prior, such truncation merely shifts rather than resolves the hallucination problem. This motivates our two-stage design, where a domain-specific deterministic skeleton serves as the structural anchor for truncated diffusion, enabling high-fidelity synthesis in $\leq$5 steps.

3 Proposed Method

Let $V_{low} \in \mathbb{R}^{d\times h\times w}$ denote the acquired anisotropic volume with high-resolution lateral dimensions and d sparse slices along the axial (Z) direction. Axial super-resolution (ASR) aims to reconstruct an axially high-resolution volume $V_{high} \in \mathbb{R}^{\hat{d}\times h\times w}$ by synthesizing intermediate slices, where the upsampling factor is $r \geq 2$ and $\hat{d} = (d-1)\times r+1$. For any two adjacent slices I_0 and I_1 in V_{low}, the objective is to generate the missing intermediate slice at a relative position $\tau \in (0, 1)$. As established by the fidelity-perception trade-off [2], directly learning an end-to-end mapping $\mathcal{F} : \{I_0, I_1, \tau\} \rightarrow \hat{I}_\tau$ is highly ill-posed, inevitably collapsing toward either regression smoothing or generative hallucinations.

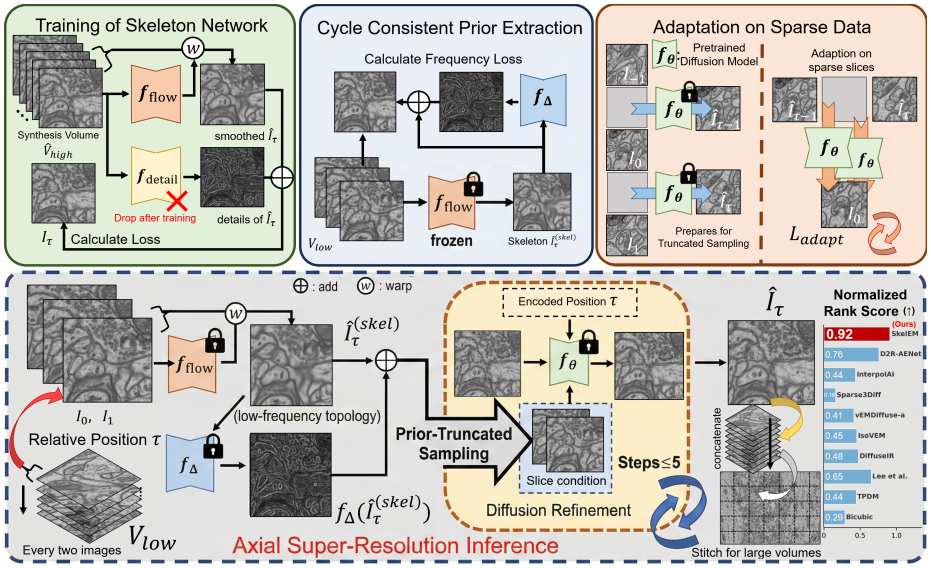

Fig. 2: Overview of the SkelEM framework. **Top: Three-stage training.** (*Left*) f_{flow} is trained on $\hat{V}_{high}$ to produce a smooth structural skeleton; f_{detail} is deliberately discarded to enforce topology–texture separation. (*Middle*) The frozen f_{flow} cyclically reconstructs observed slices from V_{low}, and a residual estimator f_Δ is trained via frequency loss to enhance the high-frequency details. (*Right*) The pretrained diffusion refiner f_θ is adapted on sparse real slices via bidirectional self-alignment loss $\mathcal{L}_{adapt}$ to eliminate synthetic bias and prepare for truncated sampling. **Bottom: Inference.** Given I_0, I_1 and target position τ, the frozen f_{flow} warps a structural skeleton $\hat{I}_\tau^{(skel)}$, and f_Δ predicts its high-frequency residual; their sum initializes the reverse diffusion at an intermediate timestep, from which f_θ completes refinement in ≤ 5 steps. The bar chart on the right summarizes the normalized rank score $\bar{s}$ across 25 metrics on FIB-25, EPFL, BRAVE-ASR (zero-shot), and EPFL membrane segmentation; SkelEM attains $\bar{s} = 0.92$, substantially ahead of all self-supervised baselines (see Suppl. for details).

SkelEM resolves this through training-signal decoupling: the topological network and the diffusion refiner are optimized by disjoint objectives with no shared gradients (illustrated in Fig. 2). Unlike cascaded pipelines that share supervision across the topology and texture stages [22, 32], this separation prevents synthetic-texture leakage into the structural skeleton while enabling each stage to specialize. First, a topological network trained on synthetic manifolds establishes a deterministic skeleton as a stable low-frequency anchor (Sec. 3.1); a diffusion refiner then recovers high-frequency biological textures, trained on synthetic manifolds and self-aligned on sparse real slices (Sec. 3.2). A lightweight residual estimator bridges the two via prior-truncated sampling (Sec. 3.3).

3.1 Topological Skeleton via Synthetic Manifold Pre-training

Biological ultrastructures exhibit highly non-linear topological deformations across the large axial gaps of V_{low}. Although flow-based VFI networks provide a natural

structural prior for slice synthesis, models pre-trained on natural videos struggle to capture the complex nonlinear dynamics of biological tissues, as we demonstrate in Sec. 4.4. This domain gap motivates training a flow network on synthetic microscopy manifolds rather than directly using natural video priors [14, 18].

To this end, we leverage a prior distillation method [4] to generate a dense pseudo-HR volume $\hat{V}_{high}$ as training data, and adopt the intermediate flow network (IFNet) from RIFE [13] as our topological network f_{flow}. Given adjacent authentic slices I_0 and I_1 at a relative position $\tau \in (0, 1)$, f_{flow} predicts bidirectional flows $F_{\tau\to0}, F_{\tau\to1}$ and a blending mask $M_\tau \in [0, 1]$. The structural skeleton $\hat{I}_\tau^{(skel)}$ is synthesized via backward warping $\mathcal{W}$:

$$\hat{I}_\tau^{(skel)} = M_\tau \odot \mathcal{W}(I_0, F_{\tau\to0}) + (1 - M_\tau) \odot \mathcal{W}(I_1, F_{\tau\to1}) \tag{1}$$

As discussed in Sec. 2, standard VFI pipelines append a detail refiner to inject high-frequency textures. Since $\hat{V}_{high}$ inherently contains cross-plane artifacts, however, retaining this refiner would inevitably leak synthetic hallucinations into the skeleton. We therefore deliberately discard the detail refiner after training on $\hat{V}_{high}$ and strictly freeze f_{flow}, restricting it to function exclusively as a deterministic low-frequency topological anchor. This enforces the central separation principle of our framework: f_{flow}'s gradients never reach the refinement stage, isolating spatial warping from the subsequent real-domain texture recovery described in Sec. 3.2. The quality and stability of $\hat{V}_{high}$ as a training signal is empirically corroborated by D2R-AENet [4], which trains exclusively on the same synthetic volume and consistently achieves the highest fidelity scores among all self-supervised baselines (see Tab. 1).

3.2 Diffusion Detail Refiner and Sparse Self-Alignment

While the topological skeleton $\hat{I}_\tau^{(skel)}$ establishes a robust low-frequency foundation, it inherently lacks the fine-grained biological textures lost during physical sectioning. To restore these missing high-frequency details, we employ a dedicated diffusion-based refinement network f_θ that learns to enhance the intermediate high-frequency distribution conditioned on boundary slices I_0 and I_1.

Base Training on Synthetic Manifolds. We first establish a base texture generation capability by training f_θ on the pseudo-HR volume $\hat{V}_{high}$ from Sec. 3.1. For an intermediate target slice $\hat{I}_\tau$ sampled from the synthetic volume, the model is optimized via the standard diffusion objective:

$$\mathcal{L}_{base} = \mathbb{E}_{t,\hat{I}_\tau,\epsilon}\left[\|\epsilon - f_\theta(x_t, t, I_0, I_1, \tau)\|_2^2\right], \tag{2}$$

where t is the diffusion timestep and x_t represents the noisy state of $\hat{I}_\tau$. This phase equips the model with a strong generative prior capturing the macroscopic biological score function without relying on massive external pre-trained models. However, since $\hat{V}_{high}$ inherently carries cross-plane artifacts, the refiner trained solely on synthetic data inevitably inherits these biases, motivating the subsequent sparse self-alignment stage.

Sparse Self-Alignment on Real Slices. To eliminate the synthetic bias accumulated during base training, we fine-tune f_θ directly on the authentic sparse slices of V_{low}. Given a sliding window of three slices $\{I_{-1}, I_0, I_1\}$, we first synthesize a pair of intermediate anchors using the current refiner: $\hat{I}_{\tau-1}$ (generated between I_{-1} and I_0) and $\hat{I}_\tau$ (generated between I_0 and I_1). Unlike Sparse3Diff [20], which relies on external pre-trained models and applies only a unidirectional constraint, we optimize f_θ via a unified *bidirectional* self-alignment objective. By alternating the conditioning sequence and its complementary geometric index, the model is required to reconstruct the central observed slice I_0 consistently regardless of synthesis direction, enforcing topological symmetry and suppressing directional bias:

$$\mathcal{L}_{adapt} = \sum_{\phi \in \Phi} \mathbb{E}_{\epsilon, I_0, t} \left[\|\epsilon - f_\theta(x_t, t, \phi)\|_2^2 \right], \tag{3}$$

where $\Phi = \{(\hat{I}_{\tau-1}, \hat{I}_\tau, 1-\tau), (\hat{I}_\tau, \hat{I}_{\tau-1}, \tau)\}$, and the synthesized anchors $\hat{I}_{\tau-1}$, $\hat{I}_\tau$ are treated as fixed conditioning inputs without gradient backward during the optimization of $\mathcal{L}_{adapt}$. Intuitively, a direction-agnostic refiner should reconstruct I_0 equally well from either temporal direction.

3.3 Fast Inference via Prior-Truncated Sampling

Standard diffusion requires hundreds of steps from Gaussian noise, rendering it computationally prohibitive for gigapixel biological volumes. To achieve highly accelerated inference without compromising fidelity, we propose a prior-truncated sampling strategy enabled by a real-domain high-frequency residual prior.

Cycle-Consistent Prior Extraction. Building upon the topological symmetry established in Sec. 3.2, we formulate a deterministic cyclic reconstruction process utilizing the frozen f_{flow}. Operating within the slice sequence $\{I_{-1}, I_0, I_1\}$ from V_{low}, we warp two virtual anchors at symmetric topological positions: $I_{-\tau}$ and $I_{1-\tau}$. The central slice is cyclically reconstructed via

$$\tilde{I}_0 = \mathcal{W}_{cycle}(I_{-\tau}, I_{1-\tau}, f_{flow}, \tau). \tag{4}$$

Their difference $\Delta_{gt} = I_0 - \tilde{I}_0$ explicitly isolates the high-frequency details lost during nonlinear warping. To generalize this extraction to unobserved slices, a lightweight estimator f_Δ is trained to predict this residual directly from the structural skeleton. To emphasize high-frequency fidelity, the estimator is supervised by the frequency L1 loss:

$$\mathcal{L}_{freq} = \|\mathcal{F}_{fft}(f_\Delta(\hat{I}_\tau^{(skel)})) - \mathcal{F}_{fft}(\Delta_{gt})\|_1, \tag{5}$$

where $\mathcal{F}_{fft}(\cdot)$ denotes the fast Fourier transform.

Initialization via Prior Truncation. As established in Sec. 2, truncated diffusion strategies are effective only when initialized from a structurally reliable prior. The cycle-consistent residual $\Delta_\tau = f_\Delta(\hat{I}_\tau^{(skel)})$ extracted above provides precisely such an anchor. For an unobserved target slice at relative position

τ, we aggregate the deterministic skeleton $\hat{I}_\tau^{(skel)}$ and Δ_τ into a prior state $\hat{I}_\tau^{(prior)} = \hat{I}_\tau^{(skel)} + \Delta_\tau$, where the predicted residual belongs to pixel space rather than noise space. Rather than initiating the reverse process from a random state $x_T \sim \mathcal{N}(0, I)$, we truncate the generative trajectory to an intermediate timestep t_S ($S \ll T$), formulating the initial latent state as:

$$x_{t_S} = \sqrt{\bar{\alpha}_{t_S}} \left(\hat{I}_\tau^{(skel)} + \Delta_\tau \right) + \sqrt{1 - \bar{\alpha}_{t_S}} \epsilon, \tag{6}$$

where $\bar{\alpha}_{t_S}$ dictates the noise schedule and $\epsilon \sim \mathcal{N}(0, I)$ is standard Gaussian noise. This forward-simulated initialization anchors the reverse trajectory to a structurally correct manifold already enriched with authentic high-frequency cues, substantially reducing the generative burden. We acknowledge that this initialization is a heuristic approximation rather than an exact inversion, and validate its effectiveness through ablation studies in Sec. 4.4.

Rapid Iterative Refinement. Starting from the prior-injected state x_{t_S}, we perform S steps of reverse sampling ($S \leq 5$) using the bidirectionally aligned refiner f_θ from Sec. 3.2. Defining a sequence of diffusion timesteps $\{t_S, t_{S-1}, \ldots, t_0 = 0\}$, for each step n traversing from S down to 1, the denoising model predicts the clean state $\hat{x}_0$ conditioned on the authentic boundary slices I_0, I_1 and the spatial index τ:

$$\hat{x}_0 = \frac{1}{\sqrt{\bar{\alpha}_{t_n}}} \left(x_{t_n} - \sqrt{1 - \bar{\alpha}_{t_n}} f_\theta(x_{t_n}, t_n, I_0, I_1, \tau) \right). \tag{7}$$

The subsequent latent state $x_{t_{n-1}}$ is then computed following the standard DDPM reverse formulation:

$$x_{t_{n-1}} = \frac{\sqrt{\alpha_{t_n}}(1 - \bar{\alpha}_{t_{n-1}})}{1 - \bar{\alpha}_{t_n}} x_{t_n} + \frac{\sqrt{\bar{\alpha}_{t_{n-1}}}(1 - \alpha_{t_n})}{1 - \bar{\alpha}_{t_n}} \hat{x}_0 + \sigma_{t_n} z. \tag{8}$$

4 Experiments

4.1 Datasets and Implementation Details

Datasets and Degradation Process To comprehensively evaluate our framework, we utilize three public volume microscopy datasets (FIB-25 [38], EPFL [1], and a Zebrafish retina VLM dataset [41]) alongside our newly proposed BRAVE-ASR dataset. While FIB-25 and EPFL (acquired via FIB-SEM) serve as our primary training and in-domain testing grounds, we introduce Benchmarking Anisotropic Volume-microscopy Evaluation for Axial Super-Resolution (BRAVE-ASR, acquired via Plasma-FIB) specifically to establish a standardized benchmark for zero-shot instrument transfer (further detailed in Sec. 4.2). To simulate the anisotropic degradation, we adopt the z-axis downsampling strategy from vEMDiffuse [27] with a factor of $r = 8$. This strategy retains the intrinsic physical noise of the original sections, faithfully simulating the anisotropic acquisition process where axial slices are physically sectioned at sparse intervals.

All datasets are split into train/validation/test sets following a 70%/15%/15% ratio. Comprehensive details regarding specimens, spatial resolutions, and the quantitative validation of the above downsampling strategy via BRAVE-ASR dataset are provided in the Supplementary Materials.

Training Details All networks were trained and evaluated on patches of 256×256 pixels by using the ADAM optimizer ($\beta_1 = 0.9, \beta_2 = 0.99$). The skeleton network was trained on $\hat{V}_{high}$ for 300 epochs by L1 loss, requiring approximately 12 hours on a single GPU. The diffusion refiner underwent two-phase training: (1) pre-training on $\hat{V}_{high}$ for 880 epochs, followed by (2) self-alignment fine-tuning ($\mathcal{L}_{\text{adapt}}$) on the sparse V_{low} for 20 epochs. We use a ResNet-9 as the detail estimator and train it for 5 epochs. All components took approximately 72 hours using two GPUs for full training. The experiments were conducted using PyTorch 2.4.1 on a Linux server equipped with three Nvidia 4090 GPUs.

Comparison with State-of-the-art Methods We benchmark our method against bicubic interpolation and a comprehensive suite of state-of-the-art self-supervised ASR methods, with supervised methods (SRUNet [11], vEMDiffuse-i [27]) included for reference. Performance is evaluated using 3D-PSNR, SSIM, and LPIPS [46]. All methods were trained following their official implementations, and evaluated on 256^2 patches following standard practice [22, 23, 27, 32].

4.2 Quantitative Performance Comparison

We applied SkelEM and comparative methods to reconstruct volumes from the FIB-25 and EPFL test datasets, calculating image similarity metrics in all three views, as shown in Tab. 1. The results directly reflect the trilemma outlined in Sec. 1: self-supervised methods are broadly trapped by a fidelity-versus-perception trade-off, with no single baseline resolving all three axes simultaneously. Regression-based networks like D2R-AENet achieve the highest fidelity scores at the cost of severe perceptual degradation, confirming their over-smoothing tendency. Other diffusion-based methods either compromise 3D structural consistency or overall fidelity under self-supervised constraints. In contrast, SkelEM resolves this trilemma through training-signal decoupling between its two specialized stages: it consistently achieves the best perceptual quality in the XZ/YZ views, secures highly competitive fidelity scores across both datasets, and as demonstrated in Tab. 5, requires merely 3 inference steps—over an order of magnitude faster than existing diffusion-based ASR methods. This unique balance across all three axes validates that dedicated separation of structural continuity and high-frequency detail synthesis is essential for practical, high-fidelity isotropic reconstruction.

Zero-Shot Instrument Transfer Comparison Generalization across disparate microscopy hardware remains a fundamental bottleneck in microscopy. To

Table 1: Quantitative evaluation on FIB25 and EPFL datasets. Best/second-best results in self-supervised methods are marked in **bold**/underlined, respectively. Supervised (Sup.) methods are listed for reference only and not included in the ranking.

Dataset	Type	Methods	3D-PSNR(↑)	SSIM(↑)			LPIPS(↓)		
				XY	XZ	YZ	XY	XZ	YZ
FIB25	Sup.	SRUNet [11]	26.71	0.7168	0.7119	0.7061	0.3718	0.4371	0.4525
		vEMDiffuse-i [27]	25.68	0.6669	0.6287	0.6210	0.2653	0.3994	0.4380
	Self-sup.	Bicubic	25.33	0.6528	0.6339	0.6254	**0.2578**	0.5881	0.5971
		TPDM [23]	23.38	0.5129	0.5186	0.5109	0.3613	0.4448	0.4203
		Lee et al. [22]	26.18	0.6607	0.6529	0.6456	0.3332	0.3862	0.3760
		DiffuseIR [32]	25.13	0.5820	0.5724	0.5663	0.4131	0.4394	0.4382
		IsoVEM [10]	24.57	0.6350	0.6329	0.6226	0.2654	0.5296	0.5578
		vEMDiffuse-a [27]	23.51	0.5699	0.5244	0.5164	0.3455	0.4260	0.4466
		Sparse3Diff [20]	22.83	0.5568	0.4131	0.4042	0.3265	0.5664	0.6056
		InterpolAI [18]	24.27	0.6140	0.5676	0.5613	0.2943	0.5535	0.5819
		D2R-AENet [4]	**27.62**	**0.7373**	**0.7271**	**0.7206**	0.3556	0.4136	0.4246
		SkelEM (ours)	26.24	0.6765	0.6547	0.6478	0.2721	**0.3658**	**0.3687**
EPFL	Sup.	SRUNet [11]	25.61	0.6476	0.6863	0.6813	0.5272	0.4678	0.4717
		vEMDiffuse-i [27]	24.45	0.5514	0.5833	0.5779	0.2652	0.3034	0.3094
	Self-sup.	Bicubic	23.09	0.4938	0.5191	0.5100	0.3012	0.6794	0.6674
		TPDM [23]	22.82	0.4976	0.5279	0.5210	**0.2859**	0.4070	0.4029
		Lee et al. [22]	24.99	0.5593	0.6044	0.5974	0.3428	0.3327	0.3431
		DiffuseIR [32]	24.48	0.5059	0.5625	0.5560	0.3982	0.3649	0.3788
		IsoVEM [10]	22.98	0.5085	0.5542	0.5436	0.3012	0.6434	0.6528
		vEMDiffuse-a [27]	22.98	0.4879	0.5092	0.5007	0.3808	0.4269	0.4060
		Sparse3Diff [20]	22.05	0.4723	0.4516	0.4438	0.3752	0.6022	0.6063
		InterpolAI [18]	23.67	0.5116	0.5426	0.5376	0.3491	0.5727	0.5861
		D2R-AENet [4]	**26.35**	**0.6387**	**0.6793**	**0.6734**	0.4764	0.4788	0.4795
		SkelEM (ours)	25.56	0.6192	0.6532	0.6464	0.3049	**0.3213**	**0.3298**

address this, a key contribution of our work is establishing a standardized zero-shot instrument transfer benchmark using the proposed BRAVE-ASR dataset. Specifically, we evaluate models trained on EPFL (Zeiss NVision40 FIB-SEM) directly on the unseen BRAVE-ASR testbed (Thermo Scientific Helios Hydra Plasma FIB-SEM). As shown in Fig. 1 and Tab. 2, existing SOTA methods struggle significantly with this domain shift. VFI-based methods such as InterpolAI [18] suffer from a natural-to-biological domain gap; while reusing input textures preserves surface-level perception, it introduces severe structural distortions and degrades axial fidelity. Pure diffusion models achieve higher perception but fail on fidelity metrics. Conversely, 3D convolutional networks like D2R-AENet [4] preserve high fidelity but produce overly smooth results with degraded perceptual quality across all planes. In contrast, SkelEM achieves the best perceptual quality in the XZ/YZ views while maintaining competitive fidelity scores across all metrics, demonstrating a favorable balance under zero-

Table 2: Quantitative evaluation of zero-shot domain transfer on the BRAVE-ASR dataset. All models were trained on the EPFL dataset and evaluated directly on BRAVE-ASR without fine-tuning. Best/second-best results in self-supervised (Self-sup.) methods are marked in **bold**/underlined, respectively. Supervised (Sup.) methods are listed for reference and not included in the ranking.

Dataset	Type	Methods	3D-PSNR(↑)	SSIM(↑)			LPIPS(↓)		
				XY	XZ	YZ	XY	XZ	YZ
BRAVE-ASR	Sup.	SRUNet [11]	23.67	0.5979	0.6274	0.6325	0.5210	0.4923	0.4810
		vEMDiffuse-i [27]	21.99	0.4777	0.4902	0.4965	0.3920	0.4257	0.3943
	Self-sup.	Bicubic	20.44	0.4352	0.4069	0.4147	0.3726	0.7236	0.7166
		TPDM [23]	21.02	0.4489	0.4598	0.4668	0.3769	0.4570	0.4203
		Lee et al. [22]	21.12	0.3998	0.4291	0.4354	0.3926	0.3634	0.3492
		DiffuseIR [32]	22.40	0.4661	0.4584	0.4673	0.4404	0.4177	0.4076
		IsoVEM [10]	23.23	0.5318	0.5571	0.5612	0.4409	0.6009	0.5894
		vEMDiffuse-a [27]	21.73	0.4729	0.4825	0.4893	0.4082	0.4346	0.3853
		Sparse3Diff [20]	21.35	0.4677	0.4451	0.4498	0.4027	0.5883	0.5559
		InterpolAI [18]	21.58	0.4472	0.4658	0.4728	**0.3695**	0.6475	0.6337
		D2R-AENet [4]	**24.15**	**0.5814**	**0.6116**	**0.6124**	0.4802	0.5351	0.5109
		SkelEM (ours)	23.48	0.5665	0.5892	0.5935	0.3747	**0.3203**	**0.3093**

shot domain shift. This balance establishes a practical foundation for deploying pre-trained SkelEM to novel imaging devices.

Segmentation Performance Comparison To assess the practical utility of our reconstructions, we evaluate downstream membrane segmentation performance on the EPFL test set. A public pre-trained segmentation model [47] is applied without fine-tuning to all reconstructed volumes, with results binarized via Otsu thresholding [31]. Qualitative comparison is shown in Fig. 3, where SkelEM produces membrane segmentations most consistent with those derived from the ground-truth volume. We report F1 score, Intersection over Union (IoU) [48], Adapted Rand Error (ARE) [24], and Variation of Information (VoI) [29], using segmentation on the ground-truth volume as reference. As shown in Tab. 3, SkelEM achieves the best performance across all four metrics among self-supervised methods, even surpassing both supervised baselines. This is particularly notable given that SkelEM does not lead on PSNR or SSIM, demonstrating that pixel-wise fidelity metrics alone are insufficient proxies for structural quality in downstream analysis.

The ranking further validates our design choices. D2R-AENet, the strongest fidelity-oriented baseline, ranks second, confirming that structural fidelity contributes to segmentation accuracy. However, SkelEM's margin over D2R-AENet across all four metrics demonstrates that recovering realistic high-frequency membrane textures is critical for precise boundary delineation. The substantial drop from vEMDiffuse-i to vEMDiffuse-a further highlights the difficulty of maintaining structural accuracy under self-supervised constraints, which is a challenge our training-signal decoupling directly addresses.

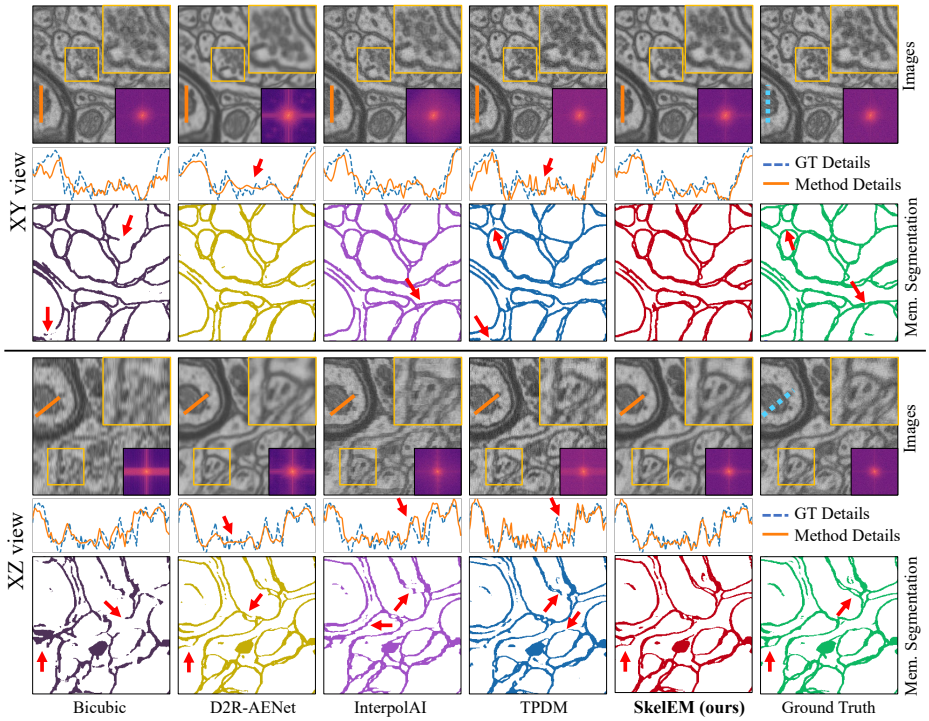

Fig. 3: Qualitative comparison of SkelEM with state-of-the-art self-supervised ASR methods on the EPFL dataset ($r = 8$) across two orthogonal views (XY and XZ). For each view, we show the reconstructed slices with zoomed-in detail insets, frequency spectra, and intensity profiles (orange) compared against the ground-truth (blue dashed), alongside the corresponding 3D membrane segmentation results. Red arrows highlight topological errors such as broken membranes and spurious connections. SkelEM (ours) best preserves high-frequency structural details, yielding membrane segmentations and details most consistent with the ground-truth among all others.

4.3 Generalizability to Volume Light Microscopy

To demonstrate generalization beyond VEM, we evaluate SkelEM on a real-world 10× ASR task using a public zebrafish retina VLM dataset [41]. Since isotropic ground truth is physically unattainable in standard VLM acquisition, quantitative evaluation is inherently infeasible in this setting, making visual assessment the standard practice for such real-world demonstrations [4, 22]. As shown in Fig. 1 (Bottom row, blue box), SkelEM recovers fine structural details from coarse anisotropic input, producing coherent 3D volumes without any supervised fine-tuning. This qualitative demonstration suggests that the training-signal decoupling design generalizes across fundamentally distinct imaging modalities.

4.4 Ablation Study

We validate the three core design axes of SkelEM on the EPFL dataset ($r = 8$), with results reported across all three views in Tab. 4.

Table 3: Membrane segmentation accuracy on the EPFL test set ($r = 8$). Best/second-best results in self-supervised methods are marked in **bold**/underlined, respectively. Supervised (Sup.) methods are listed for reference and not included in the ranking.

Type	Method	Membrane Segmentation			
		F1 Score(↑)	IoU (↑)	ARE (↓)	VoI (↓)
Sup.	SRUNet [11]	0.8233	0.6996	0.2707	0.5667
	vEMDiffuse-i [27]	0.8610	0.7560	0.2179	0.4896
Self-sup.	Bicubic	0.6969	0.5348	0.4380	0.7839
	TPDM [23]	0.8391	0.7228	0.2511	0.5464
	Lee et al. [22]	0.8319	0.7121	0.2568	0.5389
	DiffuseIR [32]	0.7885	0.6508	0.3139	0.6078
	IsoVEM [10]	0.7926	0.6564	0.3172	0.6463
	vEMDiffuse-a [27]	0.8015	0.6687	0.3035	0.6229
	Sparse3Diff [20]	0.6829	0.5185	0.4645	0.8434
	InterpolAI [18]	0.8058	0.6747	0.2962	0.6080
	D2R-AENet [4]	0.8564	0.7489	0.2240	0.4974
	SkelEM (ours)	**0.8643**	**0.7611**	**0.2129**	**0.4811**

Necessity of the two-stage design. The skeleton-only variant achieves the highest fidelity but suffers from severe axial blurring, confirming the over-smoothing tendency of convolutional priors. At the other extreme, the refiner-only variant without adaptation exhibits catastrophic fidelity collapse with severe axial artifacts, confirming that an unconstrained diffusion model cannot synthesize topologically consistent slices without structural guidance. Further applying $\mathcal{L}_{\text{adapt}}$ to this skeleton-free refiner causes additional degradation, as corrupted anchors provide misleading alignment targets and destabilize the generative prior, which demonstrates that adaptation cannot substitute for a reliable structural anchor but rather acts as a refinement component. Only the full two-stage pipeline resolves this trilemma, maintaining competitive fidelity while delivering detailed axial reconstructions.

Quality of the skeleton prior. Fixing the refinement stage and varying the skeleton source reveals a clear hierarchy. Bilinear interpolation, lacking structural priors, substantially degrades all metrics. While pre-trained VFI models provide useful motion priors, their training on natural video motion poorly generalizes to the nonlinear deformations characteristic of biological ultrastructures. Our skeleton, re-trained on the pseudo-HR volume $\hat{V}_{\text{high}}$, achieves the best performance across all metrics, validating the necessity of domain-specific re-training.

Effect of sparse data adaptation. Within the full two-stage framework, training without adaptation produces high fidelity but poor perceptual quality, as the refiner learns to reproduce the over-smoothed pseudo-HR textures. Substituting our strategy with Sparse3Diff [20] partially recovers perceptual quality but remains inferior. Our $\mathcal{L}_{\text{adapt}}$ achieves the best perceptual quality across all views with only a marginal fidelity cost, demonstrating that aligning the refiner to sparse real slices is essential for generating realistic high-frequency details.

Table 4: Ablation study on the EPFL dataset ($r = 8$). We analyze three design axes jointly: (1) **Stage**: whether to use the full two-stage pipeline; (2) **Skeleton**: the source of the initial structural prior; (3) **Adapt**: whether to apply sparse data fine-tuning. In the *Refiner only* variant, the refiner receives the sparse observed slices I_0, I_1 directly as boundary conditions, with no structural skeleton prior provided as initialization.

Stage	Skeleton	Adapt	3D-PSNR↑	SSIM↑			LPIPS↓		
				XY	XZ	YZ	XY	XZ	YZ
(i) Necessity of the two-stage design									
Skel only	Ours	—	25.96	0.6261	0.6627	0.6561	0.3466	0.5171	0.5210
Refiner only	—	×	22.43	0.5007	0.5375	0.5308	0.3115	0.4390	0.4396
Refiner only	—	✓	20.26	0.4423	0.3533	0.3456	0.3075	0.5952	0.5980
(ii) Quality of the skeleton prior									
Two-stage	Bilinear	✓	22.55	0.4822	0.4743	0.4635	0.3504	0.4071	0.4168
Two-stage	RIFE (pretrained)	✓	25.19	0.5973	0.6275	0.6216	0.3110	0.3542	0.3561
Two-stage	InterpolAI (pretrained)	✓	24.78	0.5863	0.6057	0.6082	0.3354	0.3406	0.3419
(iii) Effect of sparse data adaptation									
Two-stage	Ours	×	25.94	0.6372	0.6716	0.6659	0.4328	0.3976	0.4019
Two-stage	Ours	+Sparse3Diff [20]	25.59	0.6245	0.6578	0.6512	0.3388	0.3640	0.3701
Two-stage	Ours	✓	25.56	0.6192	0.6532	0.6464	0.3049	0.3213	0.3298

Impact of residual injection and refinement steps. We ablate the refinement steps $S \in \{1, 3, 5\}$ corresponding to initiating the reverse diffusion at timesteps $t_S \in \{100, 300, 500\}$ of a 1000-step schedule, and the residual injection strength $\lambda \in [0, 1]$ governing our truncated sampling strategy:

$$x_{t_S} = \sqrt{\bar{\alpha}_{t_S}} \left(\hat{I}_\tau^{(skel)} + \lambda \cdot \Delta_\tau \right) + \sqrt{1 - \bar{\alpha}_{t_S}}\, \epsilon, \tag{9}$$

where Δ_τ is the residual predicted by $f_\Delta(\cdot)$ and $\epsilon \sim \mathcal{N}(0, I)$ is Gaussian noise. Results are reported in Tab. 5. At $S = 1$, the full residual prior ($\lambda = 1.0$) anchors the single denoising step to a structurally informed starting point and achieves peak fidelity, while reducing λ to 0 leads to severe structural collapse, indicating Δ_τ provides essential structural information under minimal diffusion budget. Yet regardless of λ, $S = 1$ yields poor perceptual quality, as one step cannot remove the regression bias of the prior network. At $S = 3$, additional stochastic steps inject authentic biological textures, sensitivity to λ largely vanishes, and $\lambda = 1.0$ achieves the optimal equilibrium between structural fidelity and perceptual realism. Further increasing to $S = 5$ yields no quality gain while slightly degrading fidelity, suggesting over-diffusion of the structural prior. We therefore adopt $\lambda = 1.0$, $S = 3$ as our default.

5 Conclusion

In this work, we propose SkelEM, a novel two-stage self-supervised framework with training-signal decoupling between structural skeleton generation and skeleton-guided diffusion refinement. By identifying the absence of a domain-adapted structural skeleton as the shared root cause of over-smoothing and hallucination, SkelEM establishes a deterministic topological anchor that enables effective truncated diffusion in merely $\leq$5 steps. Extensive experiments demonstrate

Table 5: Ablation on residual injection strength (λ) and refinement steps S. $\lambda = 1.0$ injects the full predicted residual Δ_τ into the initial state; $\lambda = 0.0$ reduces to skeleton-only initialization without residual injection. Per-patch inference times (256×256, single RTX 4090) are listed per block. For reference, other diffusion-based ASR methods require: vEMDiffuse-i/a [27] ∼2071 ms, Lee et al. [22] ∼2558 ms, DiffuseIR [32] ∼8503 ms, TPDM [23] ∼402788 ms per patch (see Suppl. for details).

Steps (S)	Time (ms)↓	λ	3D-PSNR↑	SSIM↑			LPIPS↓		
				XY	XZ	YZ	XY	XZ	YZ
$S = 1$	84	1.00	25.93	0.6348	0.6701	0.6636	0.3486	0.3969	0.4020
		0.50	25.11	0.5827	0.6206	0.6132	0.2759	0.2975	0.3058
		0.00	23.92	0.5113	0.5510	0.5428	0.2877	0.2997	0.3070
$S = 3$	151	1.00	25.56	0.6192	0.6532	0.6464	0.3049	0.3213	0.3298
		0.50	25.49	0.6166	0.6501	0.6433	0.3035	0.3172	0.3268
		0.00	25.51	0.6150	0.6484	0.6416	0.3034	0.3186	0.3280
$S = 5$	225	1.00	25.31	0.6107	0.6411	0.6344	0.3053	0.3275	0.3356
		0.50	25.33	0.6102	0.6406	0.6339	0.3056	0.3287	0.3366
		0.00	25.36	0.6093	0.6400	0.6334	0.3054	0.3304	0.3379

the most favorable balance across the fidelity-perception trade-off among self-supervised methods, with state-of-the-art downstream membrane segmentation performance and robust generalization across distinct imaging instruments and modalities, offering a practical solution for large-scale isotropic reconstruction in connectomics and cell biology.

Acknowledgements

This work was supported by the project of high-throughput multifunctional scanning electron microscopy analytical system.

References

1. Aurelien, L., Yunpeng, L., Carlos, B., Pascal, F.: `https://www.epfl.ch/labs/cvlab/data/data-em/` (2013), accessed: 2024-08-16
2. Blau, Y., Michaeli, T.: The perception-distortion tradeoff. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6228–6237 (2018)
3. Borisovs, V., Bossi, M., Matino, L., Marmiroli, P., Cavaletti, G.: New approaches based on serial-block face electron microscopy to investigate the peripheral nervous system. Journal of the Peripheral Nervous System **30**(2), e70019 (2025)
4. Chen, B., Zhang, Y., Lv, Y., Han, H., Chen, X.: Self-supervised axial super-resolution for volume microscopy via diffusion-guided structure distillation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 467–477. Springer (2025)
5. Chen, Q., Chen, X., Wang, C., Liu, Y., Xiong, Z., Wu, F.: Learning multimodal volumetric features for large-scale neuron tracing. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 1174–1182 (2024)

6. Deng, S., Fu, X., Xiong, Z., Chen, C., Liu, D., Chen, X., Ling, Q., Wu, F.: Isotropic reconstruction of 3d em images with unsupervised degradation learning. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part V 23. pp. 163–173. Springer (2020)
7. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
8. Elliott, A.D.: Confocal microscopy: principles and modern practices. Current protocols in cytometry **92**(1), e68 (2020)
9. Hai, Y., Wang, G., Su, T., Jiang, W., Hu, Y.: Hierarchical flow diffusion for efficient frame interpolation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 22943–22952 (2025)
10. He, J., Zhang, Y., Sun, W., Yang, G., Sun, F.: Isovem: Isotropic reconstruction for volume electron microscopy based on transformer. bioRxiv pp. 2023–11 (2023)
11. Heinrich, L., Bogovic, J.A., Saalfeld, S.: Deep learning for isotropic super-resolution from non-isotropic 3d electron microscopy. In: Medical Image Computing and Computer-Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part II 20. pp. 135–143. Springer (2017)
12. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
13. Huang, Z., Zhang, T., Heng, W., Shi, B., Zhou, S.: Real-time intermediate flow estimation for video frame interpolation. In: European Conference on Computer Vision. pp. 624–642. Springer (2022)
14. Jain, S., Watson, D., Tabellion, E., Poole, B., Kontkanen, J., et al.: Video interpolation with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7341–7351 (2024)
15. Jeong, M.W., Rhee, C.E.: Lc-mamba: Local and continuous mamba with shifted windows for frame interpolation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 17671–17681 (2025)
16. Jiang, C., Gedeon, A., Lyu, Y., Landgraf, E., Zhang, Y., Hou, X., Kondepudi, A., Chowdury, A., Lee, H., Hollon, T.: Super-resolution of biomedical volumes with 2d supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6966–6977 (2024)
17. Jiang, H., Sun, D., Jampani, V., Yang, M.H., Learned-Miller, E., Kautz, J.: Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9000–9008 (2018)
18. Joshi, S., Forjaz, A., Han, K.S., Shen, Y., Queiroga, V., Selaru, F.A., Gérard, M., Xenes, D., Matelsky, J., Wester, B., et al.: Interpolai: deep learning-based optical flow interpolation and restoration of biomedical images for improved 3d tissue mapping. Nature Methods pp. 1–12 (2025)
19. Kievits, A.J., Lane, R., Carroll, E.C., Hoogenboom, J.P.: How innovations in methodology offer new prospects for volume electron microscopy. Journal of Microscopy **287**(3), 114–137 (2022)
20. Lee, H.J., Jo, E., Lim, M., Son, Y.H., Kang, B., Nam, H., Jeong, J.H., Shin, D.H., Kam, T.E.: Sparse3diff: A diffusion framework for 3d reconstruction from sparse 2d slices in volumetric optical imaging. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 510–519. Springer (2025)

21. Lee, K., Jeong, W.K.: Reference-free isotropic 3d em reconstruction using diffusion models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 235–245. Springer (2023)
22. Lee, K., Jeong, W.K.: Reference-free axial super-resolution of 3d microscopy images using implicit neural representation with a 2d diffusion prior. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 593–602. Springer (2024)
23. Lee, S., Chung, H., Park, M., Park, J., Ryu, W.S., Ye, J.C.: Improving 3d imaging with pre-trained perpendicular 2d diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10710–10720 (2023)
24. Liu, T., Jones, C., Seyedhosseini, M., Tasdizen, T.: A modular hierarchical approach to 3d electron microscopy image segmentation. Journal of neuroscience methods **226**, 88–102 (2014)
25. Liu, Y., Xie, L., Siyao, L., Sun, W., Qiao, Y., Dong, C.: Enhanced quadratic video interpolation. In: Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16. pp. 41–56. Springer (2020)
26. Liu, Y., Wang, G., Ascoli, G.A., Zhou, J., Liu, L.: Neuron tracing from light microscopy images: automation, deep learning and bench testing. Bioinformatics **38**(24), 5329–5339 (2022)
27. Lu, C., Chen, K., Qiu, H., Chen, X., Chen, G., Qi, X., Jiang, H.: Diffusion-based deep learning method for augmenting ultrastructural imaging and volume electron microscopy. Nature Communications **15**(1), 4677 (2024)
28. Ma, H., Chen, J., Deng, Z., Sun, T., Luo, Q., Gong, H., Li, X., Long, B.: Multiscale analysis of cellular composition and morphology in intact cerebral organoids. Biology **11**(9), 1270 (2022)
29. Meilă, M.: Comparing clusterings—an information based distance. Journal of multivariate analysis **98**(5), 873–895 (2007)
30. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: Sdedit: Guided image synthesis and editing with stochastic differential equations. arXiv preprint arXiv:2108.01073 (2021)
31. Otsu, N., et al.: A threshold selection method from gray-level histograms. Automatica **11**(285-296), 23–27 (1975)
32. Pan, M., Gan, Y., Zhou, F., Liu, J., Zhang, Y., Wang, A., Zhang, S., Li, D.: Diffuseir: Diffusion models for isotropic reconstruction of 3d microscopic images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 323–332. Springer (2023)
33. Park, H., Na, M., Kim, B., Park, S., Kim, K.H., Chang, S., Ye, J.C.: Deep learning enables reference-free isotropic super-resolution for volumetric fluorescence microscopy. Nature Communications **13**(1), 3297 (2022)
34. Peddie, C.J., Genoud, C., Kreshuk, A., Meechan, K., Micheva, K.D., Narayan, K., Pape, C., Parton, R.G., Schieber, N.L., Schwab, Y., et al.: Volume electron microscopy. Nature Reviews Methods Primers **2**(1), 51 (2022)
35. Schwartz, J., Jiang, Y., Wang, Y., Aiello, A., Bhattacharya, P., Yuan, H., Mi, Z., Bassim, N., Hovden, R.: Removing stripes, scratches, and curtaining with nonrecoverable compressed sensing. Microscopy and Microanalysis **25**(3), 705–710 (2019)
36. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models (2020)
37. Sun, H., Ye, E., Lu, C., Jiang, Z., Bai, W., Ren, S., Wang, C., Zhang, J., Shi, R., Ma, L., et al.: Em generalist: A physics-driven diffusion foundation model for electron microscopy (2025)

38. Takemura, S.y., Xu, C.S., Lu, Z., Rivlin, P.K., Parag, T., Olbris, D.J., Plaza, S., Zhao, T., Katz, W.T., Umayam, L., et al.: Synaptic circuits and their variations within different columns in the visual system of drosophila. Proceedings of the National Academy of Sciences **112**(44), 13711–13716 (2015)
39. Troidl, J., Liang, Y., Beyer, J., Tavakoli, M., Danzl, J.G., Hadwiger, M., Pfister, H., Tompkin, J.: niiv: Interactive self-supervised neural implicit isotropic volume reconstruction. MICCAI Workshop on Efficient Medical AI (EMA) (2025). https://doi.org/10.1101/2024.09.07.611785
40. Weigert, M., Royer, L., Jug, F., Myers, G.: Isotropic reconstruction of 3d fluorescence microscopy images using convolutional neural networks. In: Medical Image Computing and Computer-Assisted Intervention MICCAI 2017. pp. 126–134. Springer International Publishing, Cham (2017)
41. Weigert, M., Schmidt, U., Boothe, T., Müller, A., Dibrov, A., Jain, A., Wilhelm, B., Schmidt, D., Broaddus, C., Culley, S., et al.: Content-aware image restoration: pushing the limits of fluorescence microscopy. Nature methods **15**(12), 1090–1097 (2018)
42. Wu, S., Nakao, M., Imanishi, K., Nakamura, M., Mizowaki, T., Matsuda, T.: Computed tomography slice interpolation in the longitudinal direction based on deep learning techniques: To reduce slice thickness or slice increment without dose increase. Plos one **17**(12), e0279005 (2022)
43. Xu, C.S., Hayworth, K.J., Lu, Z., Grob, P., Hassan, A.M., García-Cerdán, J.G., Niyogi, K.K., Nogales, E., Weinberg, R.J., Hess, H.F.: Enhanced fib-sem systems for large-volume 3d imaging. elife **6**, e25916 (2017)
44. Ye, S., Yin, Y., Yao, J., Nie, J., Song, Y., Gao, Y., Yu, J., Li, H., Fei, P., Zheng, W.: Axial resolution improvement of two-photon microscopy by multi-frame reconstruction and adaptive optics. Biomedical Optics Express **11**(11), 6634–6648 (2020)
45. Yue, Z., Liao, K., Loy, C.C.: Arbitrary-steps image super-resolution via diffusion inversion. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 23153–23163 (2025)
46. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
47. Zhang, Y., Guo, J., Zhai, H., Liu, J., Han, H.: Segneuron: 3d neuron instance segmentation in any em volume with a generalist model. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 589–600. Springer (2024), https://github.com/yanchaoz/SegNeuron
48. Zhou, D., Fang, J., Song, X., Guan, C., Yin, J., Dai, Y., Yang, R.: Iou loss for 2d/3d object detection. In: 2019 international conference on 3D vision (3DV). pp. 85–94. IEEE (2019)