VEMamba: Efficient Isotropic Reconstruction of Volume Electron Microscopy with Axial-Lateral Consistent Mamba

Supplementary Material

1. Introduction to Mamba

1.1. State Space Models

Continuous-Time and Discretized SSMs. SSMs [3, 4] are inspired by classical state space systems that model a continuous-time signal $x(t)$ through a latent state $h(t) \in \mathbb{R}^N$ to produce an output $y(t)$. The system dynamics are governed by a linear Ordinary Differential Equation:

$$h'(t) = Ah(t) + Bx(t), \quad y(t) = Ch(t), \tag{1}$$

where $A \in \mathbb{R}^{N\times N}$, $B \in \mathbb{R}^{N\times 1}$, and $C \in \mathbb{R}^{1\times N}$ are the state, input, and output matrices, respectively.

To be applied to discrete data x_t, this continuous system must be discretized. This is achieved by introducing a time-step parameter Δ and transforming (A, B) into discrete parameters $(\overline{A}, \overline{B})$ using a discretization rule, such as the Zero-Order Hold (ZOH):

$$\overline{A} = e^{\Delta A}, \quad \overline{B} = (\Delta A)^{-1}(e^{\Delta A} - I)\Delta B, \tag{2}$$

This results in a discrete-time formulation that can be computed as a linear recurrence:

$$h_t = \overline{A}h_{t-1} + \overline{B}x_t, \quad y_t = Ch_t. \tag{3}$$

The Convolutional Duality of LTI SSMs. A critical property of these Linear Time-Invariant (LTI) models (where A, B, C, Δ are fixed) is their dual nature. While Eq. (3) represents the system as a recurrent (RNN) model, it can also be expressed as a discrete convolution $y = x * \overline{K}$. By unrolling the recurrence, the entire output sequence can be computed in parallel by convolving the input x with a global convolutional kernel $\overline{K}$ of length L:

$$y = x * \overline{K}, \quad \overline{K} = (C\overline{B}, C\overline{AB}, \ldots, C\overline{A}^{L-1}\overline{B}) \tag{4}$$

This convolutional representation is highly efficient for parallel training, while the recurrent representation is efficient for auto-regressive inference.

Mamba: Selective State Space Models. The LTI property, while efficient, is also a limitation. The system's dynamics are fixed and independent of the input, constraining its ability to model content-dependent phenomena. Mamba addresses this by introducing a selective mechanism, making the system parameters time-varying and input-dependent.

Specifically, Mamba parameterizes the key system matrices B and C, as well as the time-step Δ, as functions of the current input x_t:

$$\Delta_t = s_\Delta(x_t), \quad B_t = s_B(x_t), \quad C_t = s_C(x_t) \tag{5}$$

where s_Δ, s_B, s_C are typically small neural networks. This transforms the LTI recurrence of Eq. (3) into a time-varying system. The discrete parameters $\overline{A}_t$ and $\overline{B}_t$ are recomputed at each step using the dynamic Δ_t and B_t (derived from Eq. (2)):

$$h_t = \overline{A}_t h_{t-1} + \overline{B}_t x_t, \quad y_t = C_t h_t \tag{6}$$

This simple modification is profoundly impactful. By making the state transitions conditional on the input, the model can selectively choose what information to propagate in its state h_t and what to forget. This input-dependent dynamic breaks the time-invariance necessary for the global convolution (Eq. (4)), but grants the model significantly enhanced expressive power. This time-varying recurrence is computed efficiently using hardware-aware parallel scan algorithms, retaining the computational benefits of SSMs while enabling content-based reasoning.

1.2. SSMs for Vision

The recent emergence of State Space Models (SSMs) [2, 12], particularly Mamba [3], has presented a compelling alternative to Transformers [13] for long-sequence modeling. Mamba's core innovation, the Selective Scan (S6) module, achieves linear-time complexity while effectively capturing long-range dependencies, addressing the quadratic complexity challenge that hinders Transformers in high-resolution tasks. This efficiency has spurred its rapid adoption across diverse computer vision domains, spanning both high-level tasks [7, 10, 16] and low-level image processing [1, 5, 15].

However, a fundamental challenge remains in adapting the inherently 1D SSM mechanism to process 2D spatial data. A common paradigm, therefore, involves flattening 2D images into multiple 1D sequences. Pioneering works like Vim [16] and VMamba [10] established this approach. Vim [16] introduced a bidirectional Mamba block, processing image patches sequentially and demonstrating significant speed and memory advantages over Vision Transformers (ViTs) at high resolutions. Concurrently, VMamba [10] proposed a "Cross-Scan" strategy, which flattens the input along four directions (row-wise and column-wise, both forward and backward) to integrate spatial context.

Building on these foundational methods, subsequent research has explored a variety of alternative scanning strategies to better align with the 2D structure of images. For instance, continuous scanning paths [14] and local four-directional scans [7] were introduced to enhance structural

continuity and local feature acquisition. Other efforts, such as EfficientVMamba [11], utilizes skip sampling to improve scanning efficiency.

Despite this rapid progress, these methods are fundamentally designed for 2D images and face critical limitations when applied to volumetric data. Existing scanning strategies often disrupt the spatial structure coherence [5, 15], a problem that is exacerbated when naively extended to 3D volumes. While some models like MambaIR [5] and UVM-Net [15] attempt to re-introduce locality using additional CNN layers, this compromises the computational efficiency inherent to the Mamba architecture. Furthermore, these manually designed, fixed scanning paths [7] lack the adaptability to model the complex, anisotropic relationships inherent in VEM data. This often results in a failure to preserve consistency between the high-resolution lateral planes and the sparsely sampled axial dimension. Our work, VEMamba, is designed to overcome these specific challenges by introducing a 3D-native scanning mechanism that explicitly enforces axial-lateral consistency.

2. Details of VEMamba

2.1. Training Strategy

The training of VEMamba is conducted through a designed two-stage strategy.

Stage 1: Degradation Representation Learning. The initial stage is dedicated to training the degradation encoder. The primary objective of this stage is to learn a robust representation of the degradation characteristics present in the anisotropic data. For this purpose, we adopt the hyperparameter configuration from the CDFormer [9]. This pretraining phase enables our model to effectively capture the complex transformations, such as blur and noise, that differentiate the low-resolution inputs from their high-resolution counterparts.

Stage 2: Isotropic Reconstruction. Upon the completion of the encoder training, we freeze its weights to preserve the learned degradation knowledge. Subsequently, the main backbone of the VEMamba model is trained end-to-end. In this stage, the model learns to perform the core task of isotropic reconstruction, guided by the total loss function $\mathcal{L}_{\text{total}}$, as formulated in the main paper. This two-stage approach ensures that the reconstruction process is explicitly conditioned on the learned degradation features, leading to a more targeted and effective restoration.

2.2. Inference Speed

We evaluate the inference speed and GPU memory consumption for different methods, with results presented in Table 1. As shown, our VEMamba demonstrates the highest memory efficiency, consuming only 3308 MiB of GPU memory, which is slightly lower than IsoVEM (3438 MiB) and significantly less than EMDiffuse (6660 MiB).

Table 1. Inference speed (volumes/s) and GPU memory usage (MiB) comparison on a NVIDIA RTX 4090 with batch size 1.

Method	GPU Memory↓	Inference Speed ↑
IsoVEM	3438	4.524
EMDiffuse	6660	0.014
Ours (VEMamba)	3308	2.295
Ours (U-Net)	2036	4.598

Although IsoVEM achieves a higher throughput (4.524 vol/s) than VEMamba (2.295 vol/s), its speed advantage mainly comes from its U-Net–based Transformer architecture, which processes image patches in parallel while progressively downsampling feature maps to reduce computation.

In contrast, VEMamba performs sequential state-space modeling directly on full-resolution features, making it inherently slower. To verify the effect of architectural choices, we also implement a U-Net version of VEMamba, which is an ablation variant and attains higher speed (4.598 vol/s) and lower memory usage (2036 MiB) at the cost of a slight performance drop. Since high reconstruction quality is critical in medical imaging, we prioritize accuracy over speed. Improving the performance of the U-Net variant will be an important direction for future work.

2.3. Loss Function Selection

The selection of our composite loss function, $\mathcal{L}_{\text{total}}$, was guided by an empirical study to determine the most effective objective for VEM isotropic reconstruction. Our final formulation is a weighted sum of the $\mathcal{L}_1$ loss and the Structural Similarity Index (SSIM) loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_1(Y, \hat{Y}) + \mathcal{L}_{\text{SSIM}}(Y, \hat{Y}) \tag{7}$$

where Y and $\hat{Y}$ represent the ground-truth and the reconstructed volumes, respectively.

To validate this choice, we conducted an ablation study comparing three distinct loss formulations: (1) a pure $\mathcal{L}_1$ loss, (2) a combination of $\mathcal{L}_1$, SSIM, and the LPIPS loss, and (3) our proposed $\mathcal{L}_1 + \mathcal{L}_{\text{SSIM}}$ combination. The quantitative results are presented in Table 2. As the results demonstrate, our selected loss function achieves superior performance in terms of both PSNR and SSIM, validating its suitability for preserving both pixel-level accuracy and structural integrity.

3. Details of MoCo

3.1. Structure Detail

In our framework, the degradation representation is extracted using an encoder module within the Momentum

Table 2. Ablation study on the loss function. Our proposed $\mathcal{L}_1 + \mathcal{L}_{\text{SSIM}}$ configuration yields the best results.

Loss Function	Metrics	
	PSNR	SSIM
L1	29.389	0.7659
L1+SSIM+LPIPS	29.364	0.7696
L1+SSIM	**29.442**	**0.7707**

Contrast [6] setup. The architectural design of this encoder is crucial for effectively capturing the nuanced features of various degradation types. This section provides a detailed breakdown of its structure.

The encoder is constructed by serially stacking eight identical residual blocks, followed by a final average pooling layer to produce the feature vector. The structure of Encoder is illustrated in Figure 1.

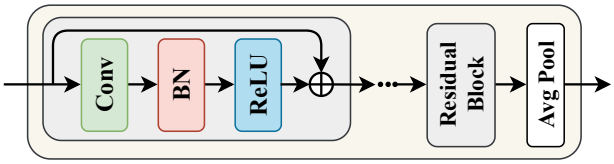

Figure 1. The structure of MoCo encoder. It comprises a convolutional layer, a Batch Normalization layer, and a ReLU activation function, integrated with a skip connection.

Specifically, within each block, the input feature map first passes through a convolutional layer (Conv). The output is then normalized using a Batch Normalization (BN) layer to stabilize the training process and accelerate convergence. Following normalization, a Rectified Linear Unit (ReLU) activation function is applied to introduce nonlinearity. A characteristic of this architecture is the residual connection, where the input to the block is added elementwise to the output of the ReLU activation. The sequential arrangement of these eight blocks allows the encoder to progressively build a rich and hierarchical representation of the input degradation patterns.

3.2. Hyperparameter Selection

To effectively simulate realistic anisotropic data for our self-supervised training paradigm, we employ a degradation pipeline that incorporates Gaussian blur, downsampling, and noise injection. The parameters governing this process are crucial for the model's ability to generalize to real-world data.

Following the methodology established in DiffuseEM [8], our primary configuration for the degradation simulation utilizes a Gaussian blur kernel with a filter size of $f = 8$ and a standard deviation of $\sigma = 4$. To substantiate this choice, we performed a comprehensive ablation study, comparing our selected parameters against several alternatives, including a baseline with only downsampling. The results, summarized in Table 3, clearly indicate that the configuration with $f = 8$ and $\sigma = 4$ yields the highest PSNR and SSIM scores. This confirms its effectiveness in generating a challenging yet representative training set for our reconstruction task.

Table 3. Ablation study on degradation simulation parameters. The setting of $f = 8, \sigma = 4$ provides the best performance.

Degradation	Metrics	
	PSNR	SSIM
Baseline (downsample)	28.997	0.7532
f=8, σ = 2	29.384	0.7628
f=8, σ = 6	29.401	0.7699
f=10, σ = 4	29.393	0.7683
f=6, σ = 4	29.389	0.7645
f=8, σ = 4	**29.442**	**0.7707**

3.3. Effectiveness of MoCo

To validate the efficacy of our self-supervised degradation learning strategy, we visualize the latent feature distributions of sub-volumes with varying degradation patterns using t-SNE. Figure 2 presents a comparison of the feature space before and after employing Momentum Contrast.

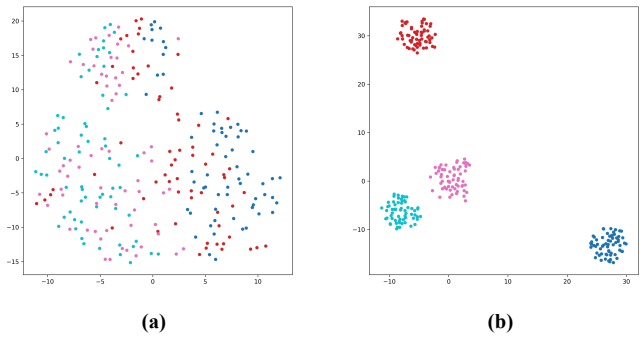

Figure 2. t-SNE visualization of degradation feature distributions. (a) Feature embedding without MoCo training, showing entangled representations where different degradation patterns are indistinguishable. (b) Feature embedding with our MoCo training strategy. The emergence of distinct clusters indicates that the encoder effectively learns to discriminate between different degradation types.

As observed in Figure 2(a), without contrastive learning, the feature representations of distinct degradation types are heavily entangled and indistinguishable. This lack of separation hinders the network's ability to perceive specific degradation levels. In stark contrast, Figure 2(b)

demonstrates that our MoCo-based training effectively disentangles these representations, resulting in clear, compact clusters corresponding to specific degradation characteristics. This distinct separation confirms that our encoder successfully learns a discriminative degradation prior. Consequently, this enables the Volume Degradation Injection Module (VDIM) to provide precise, degradation-aware modulation to the reconstruction backbone, ensuring robustness against varying anisotropic conditions.

4. Visualization of EPFL dataset

4.1. 3D Volume Visualization

To further complement the 2D slice comparisons presented in the main paper, we provide 3D volumetric visualizations of a representative sub-volume from the EPFL dataset. The following figures demonstrate the performance of our VEMamba against baseline interpolation, IsoVEM, and EMDiffuse across x4, x8, and x10 axial degradation factors.

Figure 3 showcases the 3D renderings of the reconstructed tissue volumes. As can be observed, the baseline interpolation method yields overly smooth and blurry structures, failing to recover fine details, which is consistent with its poor 2D performance. The results from IsoVEM, while improved, exhibit noticeable artifacts such as spurious, disconnected fragments and unnatural surface textures, particularly at higher magnification factors (x8, x10). This aligns with the "hallucinated boundaries" noted in the 2D axial slices (Figure 5 in the main paper). Similarly, EMDiffuse struggles to maintain structural completeness, resulting in volumes that appear eroded or contain discontinuities. In stark contrast, our VEMamba method consistently produces reconstructions with superior structural coherence and surface integrity. The membranes and organelles are rendered with remarkable clarity and continuity, preserving the complex topology of the neural tissue.

Furthermore, the quality of isotropic reconstruction is critical for the accuracy of downstream quantitative analyses. We evaluated this by performing 3D mitochondria segmentation on the reconstructed volumes, as visualized in Figure 4. The segmentations derived from the baseline, IsoVEM, and EMDiffuse reconstructions are heavily fragmented, containing numerous holes and false negatives. These artifacts would severely compromise any subsequent morphological or statistical analysis. The segmentation based on VEMamba's output, however, is significantly more complete and topologically sound. The mitochondria are rendered as continuous, well-defined objects, closely resembling the ground truth. This illustrates that the high fidelity of our reconstruction directly translates into more reliable and accurate results for downstream tasks, corroborating the superior IoU scores reported in the main paper (Table 2).

In summary, these 3D visualizations underscore the superiority of VEMamba in generating high-quality, coherent, and artifact-free isotropic volumes, which is essential for both qualitative inspection and robust quantitative analysis in biomedical research.

4.2. Axial Section Visualization

While 2D lateral fidelity is important, the definitive challenge of isotropic reconstruction lies in recovering the missing information along the undersampled axial axis. Figure 5 provides a qualitative comparison of reconstructed slices in the axial (xz and yz) planes.

The Baseline method (interpolation) exhibits severe aliasing and blurring, failing to recover any meaningful high-frequency structural details. While IsoVEM recovers sharper textures, it suffers from significant structural inconsistencies; specifically, it generates "hallucinated" boundaries that appear plausible in 2D but result in jagged, discontinuous membrane profiles when viewed axially. EMDiffuse similarly struggles with structural integrity, leading to broken organelle boundaries and noise artifacts.

Conversely, VEMamba demonstrates superior axial consistency. Our method reconstructs smooth, continuous membranes and organelles that closely align with the Ground Truth. The vertical coherence of these structures confirms the effectiveness of the Axial-Lateral Chunking Selective Scan Module (ALCSSM) in effectively modeling 3D spatial dependencies, thereby preventing the slice-to-slice discontinuity observed in competing methods.

5. Visualization of CREMI dataset

To complement the quantitative evaluation presented in Table 1 of the main paper, we provide a detailed qualitative comparison of VEMamba's reconstruction performance on the real-world anisotropic CREMI dataset. The following visualizations further substantiate the superior performance of our proposed method.

Figure 6 illustrates the reconstruction fidelity on the lateral (xy) plane at different degradation factors (x4, x8, and x10). These comparisons are performed by applying our self-supervised framework to the high-resolution lateral sections. It is immediately evident that the Baseline method (representing standard interpolation) fails to recover fine structural details. As the degradation factor increases, the baseline results become progressively blurred, and crucial ultrastructural information. This is particularly noticeable in the x8 and x10 results, where the output is smeared and lacks clarity. In stark contrast, VEMamba demonstrates exceptional performance across all factors. Our model successfully reconstructs sharp, high-frequency details, producing images that are virtually indistinguishable from the Ground Truth (GT). Even at the highly challenging x10 factor, VEMamba robustly restores crisp cell membranes and

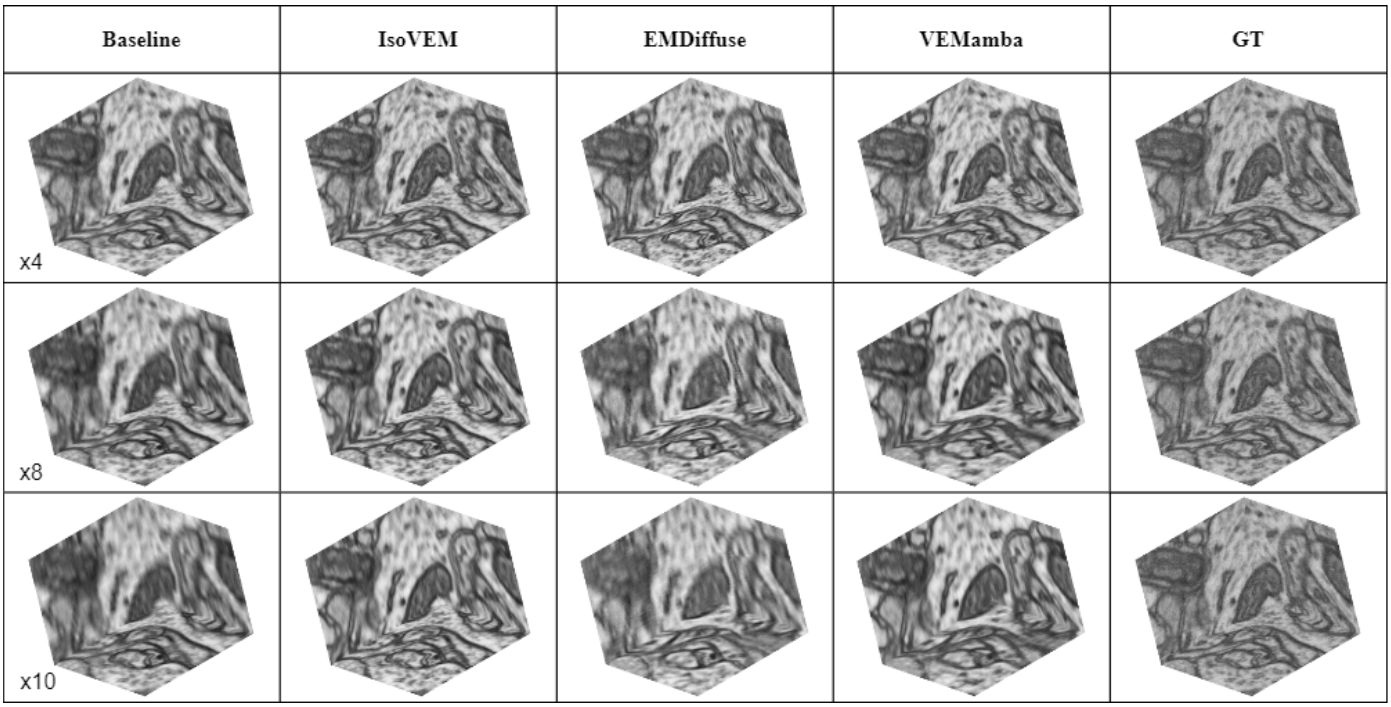

Figure 3. 3D volumetric renderings of isotropic reconstruction results on a sub-volume of the EPFL dataset. Our method consistently produces more structurally coherent and detailed reconstructions across all magnification factors (x4, x8, x10) compared to baseline interpolation and competing methods.

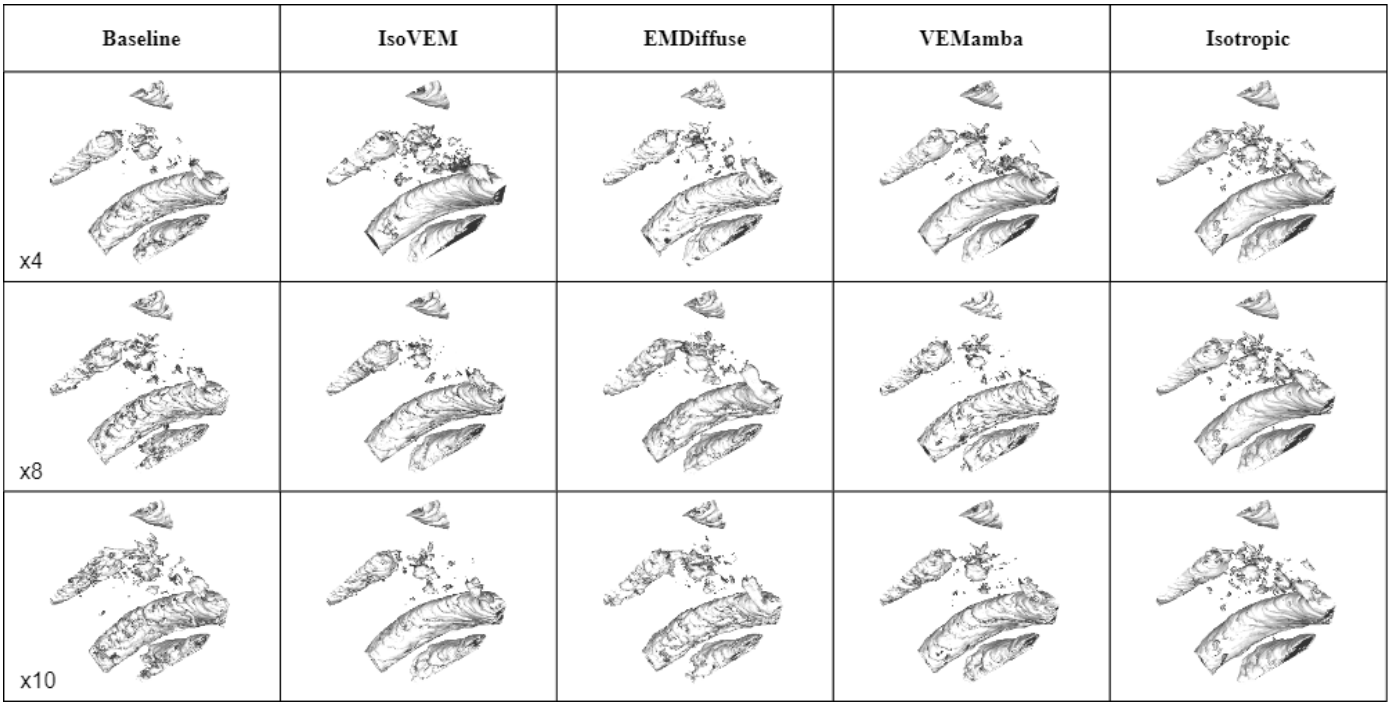

Figure 4. 3D visualization of downstream mitochondria segmentation results. The segmentations are based on the reconstructed volumes shown in Figure 3. VEMamba's output enables a significantly more complete and accurate 3D segmentation, with fewer discontinuities and artifacts, underscoring its practical utility for quantitative biological analysis.

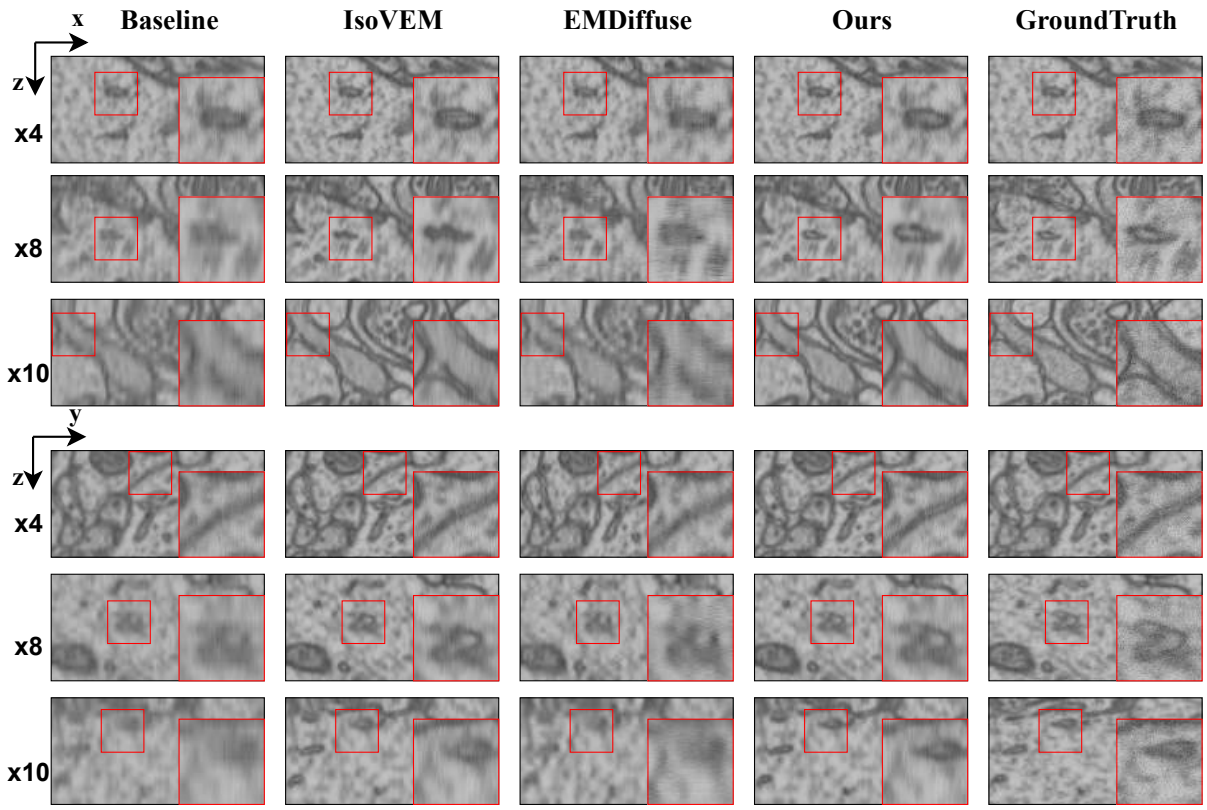

Figure 5. Visual comparison of baseline (Interpolation), IsoVEM, EMDiffuse, VEMamba, and Ground Truth on the EPFL dataset at scale factors of ×4, ×8, and ×10 in the axial (xz and yz) direction.

clearly defined organelles.

Beyond 2D fidelity, the ultimate goal of isotropic reconstruction is to achieve 3D structural consistency, which is critical for downstream analyses like segmentation and connectome tracing. Figure 7 presents 3D volume. VEMamba's 3D reconstruction exhibits outstanding spatial consistency. The rendered volume is sharp, and the surfaces of individual cells and organelles are smooth and coherent, faithfully replicating the 3D structure of the Ground Truth. This visualization powerfully demonstrates the success of our Axial-Lateral Chunking Selective Scan Module (ALCSSM) and Dynamic Weights Aggregation Module (DWAM) in capturing long-range, anisotropic spatial dependencies.

In summary, these qualitative results strongly align with our quantitative findings, confirming that VEMamba not only excels at 2D detail restoration but also achieves state-of-the-art 3D consistency, providing a reliable and high-fidelity solution for VEM isotropic reconstruction.

6. Transferability of VEMamba

To assess the generalization capability and practical utility of VEMamba, we conduct a challenging transferability study. This is crucial as real-world VEM applications often involve data from diverse sources, exhibiting significant domain shifts due to different imaging modalities, sample preparation protocols, and inherent artifact patterns. We evaluate the model's performance in two scenarios:

- **Zero-shot transfer (Ours (wo finetune))**: The model is trained on one dataset (e.g., EPFL) and directly applied to the other (e.g., CREMI) without any re-training.
- **Fine-tuned transfer (Ours (finetune))**: The model pre-trained on the source dataset is then briefly fine-tuned on a small portion of the target dataset.

The quantitative results of this cross-domain evaluation are presented in Table 4, with corresponding visual comparisons provided in Figure 8.

The results clearly demonstrate VEMamba's robust transferability. As shown in Table 4, our zero-shot model (Ours (wo finetune)) significantly outperforms the baseline across all metrics and transfer directions. For instance, in the EPFL→CREMI (x8) task, our zero-shot

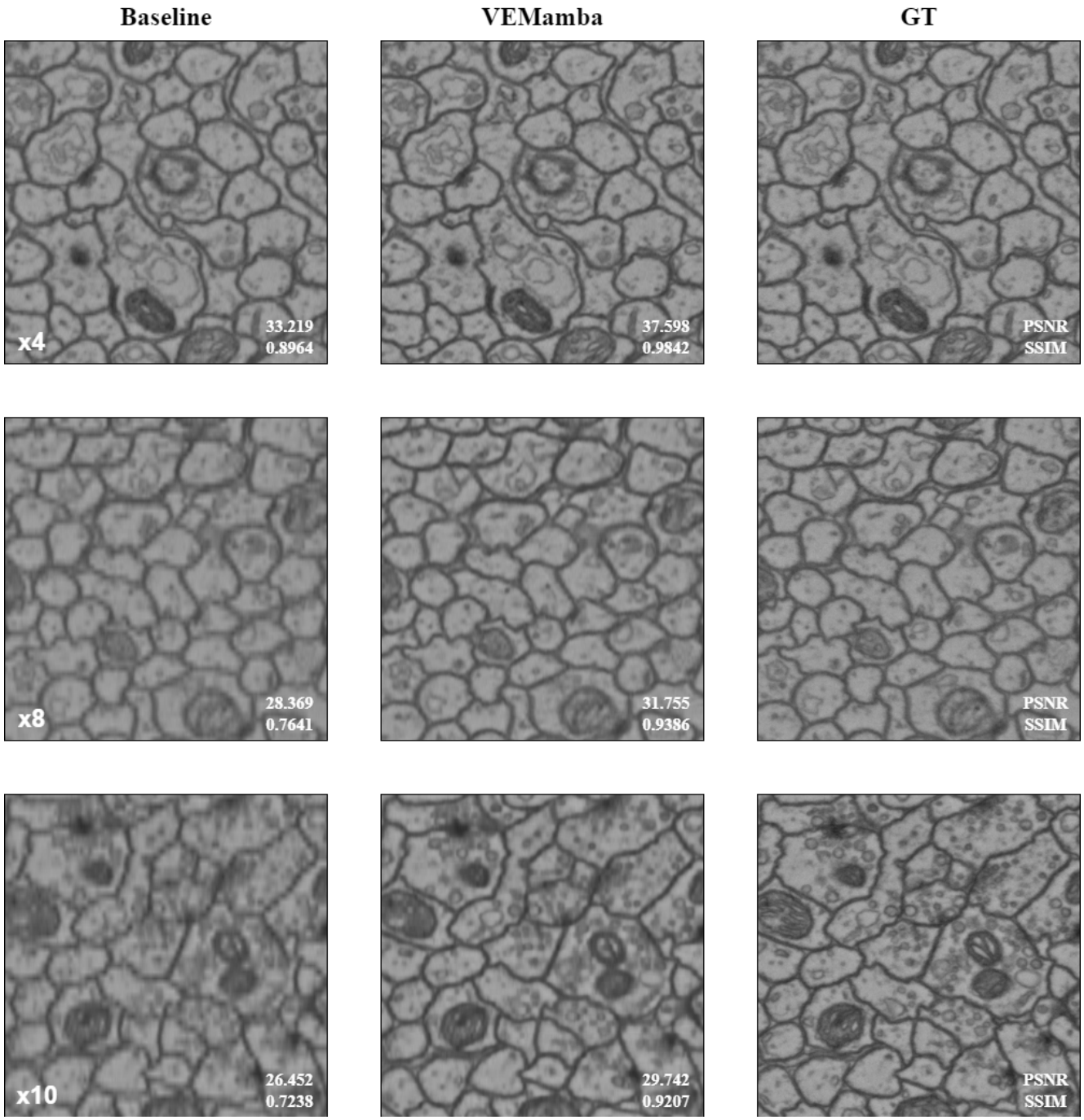

Figure 6. Qualitative comparison on the CREMI dataset (xy-plane). This figure displays 2D lateral sections reconstructed at x4, x8, and x10 degradation factors. VEMamba consistently preserves fine ultrastructural details and sharp boundaries, significantly outperforming the blurry results of the baseline method and closely matching the Ground Truth.

model achieves 30.603 PSNR, substantially surpassing the baseline's 28.129 PSNR. A similar trend is observed in the CREMI→EPFL (x8) task, where our model (27.392 PSNR) again vastly exceeds the baseline (25.558 PSNR).

Crucially, we observe that fine-tuning the model on the target domain (Ours (finetune)) yields only a **marginal improvement**. For example, in the EPFL→CREMI (x8) task, fine-tuning provides only 0.128 dB (30.731 vs. 30.603) of additional gain. This minimal gap strongly indicates that VEMamba, guided by its axial-lateral consistency

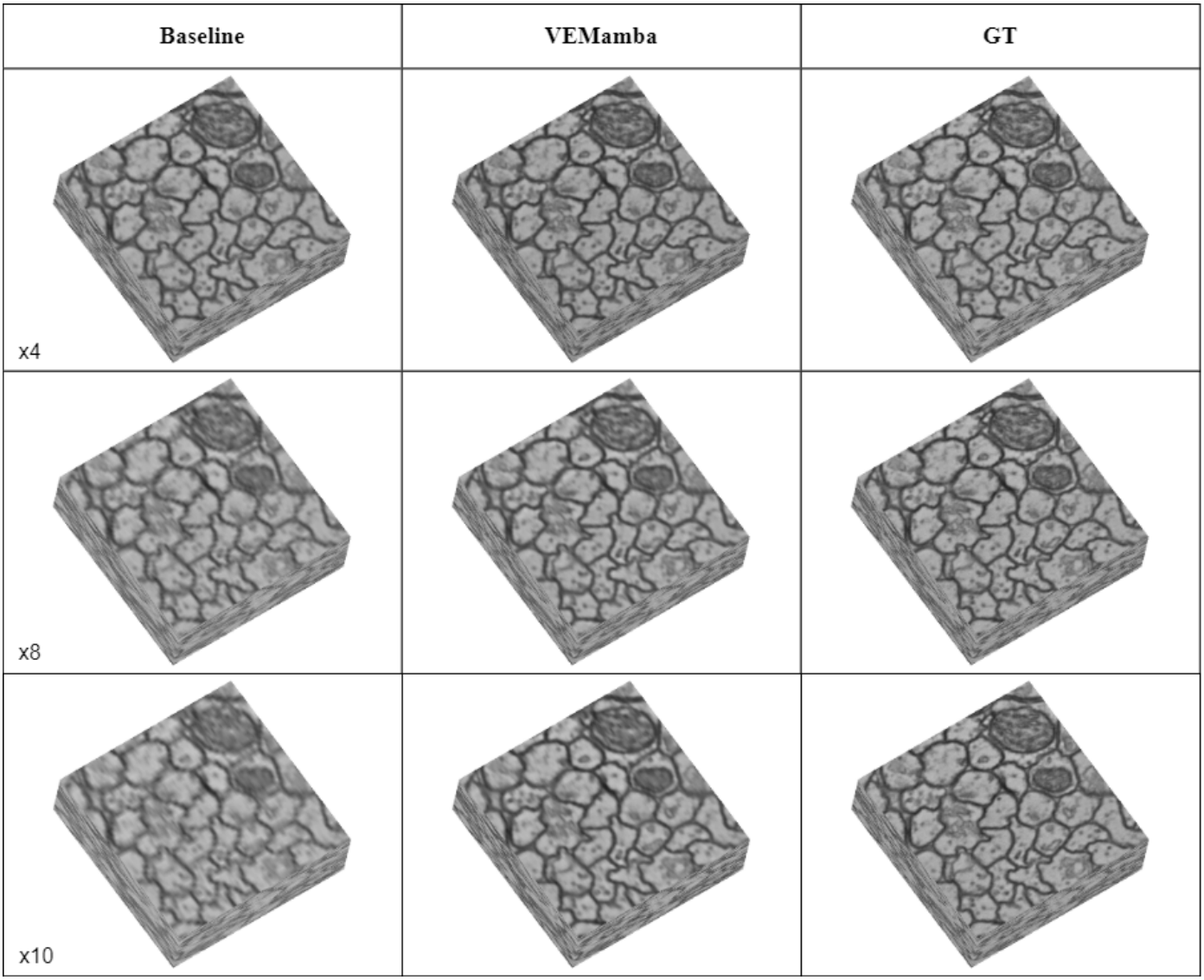

Figure 7. Qualitative comparison of 3D volume rendering on the CREMI dataset.

mechanism, learns fundamental and transferable representations of biological ultrastructures rather than overfitting to modality-specific artifacts.

This observation is visually corroborated in Figure 8. The reconstructions from our zero-shot model are markedly superior to the baseline, exhibiting significantly enhanced sharpness and structural coherence (e.g., clearer membranes). Furthermore, the visual quality of the "Ours (wo finetune)" results is virtually indistinguishable from the "Ours (finetune)" results.

This strong zero-shot performance underscores VEMamba's exceptional generalization. It suggests that our model captures the underlying principles of 3D ultrastructure, making it a robust and practical tool for real-world VEM isotropic reconstruction, even when faced with significant domain shifts.

References

[1] Rui Deng and Tianpei Gu. Cu-mamba: Selective state space models with channel learning for image restoration, 2024.

[2] Daniel Y. Fu, Tri Dao, Khaled K. Saab, Armin W. Thomas, Atri Rudra, and Christopher Ré. Hungry hungry hippos: Towards language modeling with state space models, 2023.

[3] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *arXiv preprint arXiv:2312.00752*, 2024.

[4] Albert Gu, Karan Goel, and Christopher Re. Efficiently modeling long sequences with structured state spaces. In *arXiv preprint arXiv:2111.00396*, 2022.

Table 4. Quantitative cross-domain transferability results. We compare the baseline, our zero-shot model (wo finetune), and our fine-tuned model (finetune) on tasks transferring between EPFL (FIB-SEM) and CREMI (ssTEM) datasets.

Method	Metrics	EPFL→ CREMI			CREMI→ EPFL		
		×4	×8	×10	×4	×8	×10
Baseline	PSNR	34.219	28.129	26.554	28.407	25.558	24.583
	SSIM	0.9133	0.7861	0.7379	0.7478	0.6323	0.5891
	LPIPS	0.2811	0.4445	0.5031	0.2709	0.3955	0.4398
Ours (wo finetune)	PSNR	36.804	30.603	28.932	29.345	27.392	26.438
	SSIM	0.9384	0.8432	0.8018	0.7632	0.6701	0.6355
	LPIPS	0.2089	0.3363	0.3811	0.2297	0.3239	0.3623
Ours (finetune)	PSNR	36.953	30.731	29.102	29.388	27.407	26.465
	SSIM	0.9403	0.8523	0.8147	0.7676	0.6718	0.6401
	LPIPS	0.2066	0.3237	0.3696	0.2237	0.3137	0.3542

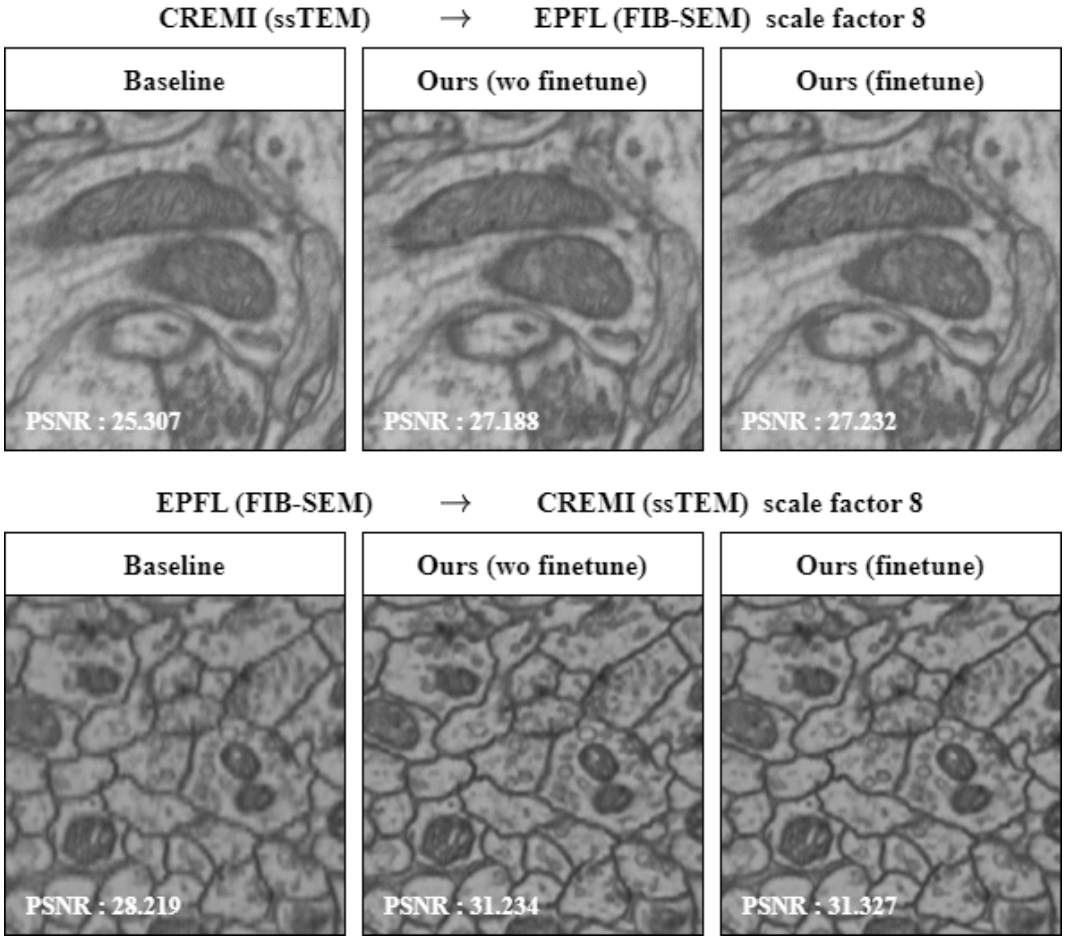

Figure 8. Visual comparison of cross-domain transferability at x8 magnification. Top: CREMI→EPFL. Bottom: EPFL→CREMI. VE-Mamba (wo finetune) achieves results comparable to the fine-tuned model and vastly superior to the baseline, demonstrating strong generalization.

[5] Hang Guo, Jinmin Li, Tao Dai, Zhihao Ouyang, Xudong Ren, and Shu-Tao Xia. Mambair: A simple baseline for image restoration with state-space model, 2024.

[6] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Confer-*

ence on Computer Vision and Pattern Recognition (CVPR), pages 9729–9738, 2020.

[7] Tao Huang, Xiaohuan Pei, Shan You, Fei Wang, Chen Qian, and Chang Xu. Localmamba: Visual state space model with windowed selective scan, 2024.

[8] Kyungryun Lee and Won-Ki Jeong. Reference-Free Isotropic 3D EM Reconstruction using Diffusion Models, 2023. arXiv:2308.01594 [cs].

[9] Qingguo Liu, Chenyi Zhuang, Pan Gao, and Jie Qin. Cd-former: When degradation prediction embraces diffusion model for blind image super-resolution. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7455–7464, 2024.

[10] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. Vmamba: Visual state space model, 2024.

[11] Xiaohuan Pei, Tao Huang, and Chang Xu. Efficientvmamba: Atrous selective scan for light weight visual mamba, 2024.

[12] Jimmy T. H. Smith, Andrew Warrington, and Scott W. Linderman. Simplified state space layers for sequence modeling, 2023.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[14] Chenhongyi Yang, Zehui Chen, Miguel Espinosa, Linus Ericsson, Zhenyu Wang, Jiaming Liu, and Elliot J. Crowley. Plainmamba: Improving non-hierarchical mamba in visual recognition, 2024.

[15] Zhuoran Zheng and Chen Wu. U-shaped vision mamba for single image dehazing, 2024.

[16] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model, 2024.