Three-Step Conditional Diffusion 3D Reconstruction for Light-Field Microscopy

Supplementary Material

1. Supplementary Material

This document serves as the **supplementary material for CVPR 2026 paper #44199**, titled *"Three-Step Conditional Diffusion 3D Reconstruction for Light-Field Microscopy."*

The purpose of this supplement is to provide additional experimental evidence, technical parameters, and visual results that support the primary claims in the main manuscript. The contents are organized as follows:

- **Experimental Details:** In-depth description of dataset generation, optical parameters, and implementation settings.
- **Comparative Visualizations:** 3D reconstruction results based on DDPM and extensive visual comparisons between the proposed TCD and current State-of-the-Art (SOTA) methods.
- **Robustness Analysis:** Evaluation of TCD and baseline methods under varied noise levels to demonstrate reliability in practical microscopy.
- **Generalization Study:** Visualizations of the mixed training dataset and performance analysis of distributional adaptation.
- **Ablation of Fine-tuning:** Training results and reconstruction performance of the fine-tuning process guided by the ICD module.

2. Experimental and Implementation Details

2.1. Dataset Preparation and Generation

Our experimental dataset comprises five distinct biological categories designed to evaluate the reconstruction fidelity and generalization of the TCD framework. To balance structural complexity and authenticity, we combine synthetic data with real-world captured 3D volumes. Specifically, the dataset includes synthetic volumes (*Tubulin*, *Mitochondria*, and *Bead cell*) generated using structurally accurate simulation models, as well as real biological specimens (*Dendrite* and *Vessel*) captured via high-resolution wide-field cameras. This combination provides a controlled yet representative variety of sample morphologies.

To enhance the model's generalization capability, data augmentation was applied during the training phase, including random rotations and multi-scale scaling of the 3D volumes and their corresponding 2D light-field images. These operations effectively improve the model's robustness to diverse spatial distributions and sample scales.

The GT volumes are standardized to a spatial dimension of $D \times H \times W = 61 \times 176 \times 176$. The axial depth $D = 61$ corresponds to a physical range of $z \in [-30, 30]$ μm with a step size of 1 μm. The lateral dimensions are normalized to 176×176 pixels to align with the microlens array sampling grid. Intensity values of all data are normalized to the range $[-1, 1]$.

2.2. Forward Projection and Optical Configuration

We employ a wave-optics-based forward projection model to generate 2D light-field observations. All synthesized light-field images utilize physical parameters derived from a real-world light-field camera setup to simulate practical imaging environments.

The detailed optical configuration is as follows: a $40\times$ objective lens with a numerical aperture (NA) of 0.8 is used in a water-immersion medium ($n = 1.33$), with an emission wavelength of $\lambda = 580$ nm. The microlens array (MLA) features a pitch of 150 μm and a focal length of 3500 μm. The system extracts $11 \times 11 = 121$ angular views at each spatial location, achieving a discretized sampling of both spatial and angular information.

2.3. Training and Implementation Details

The TCD framework is implemented in PyTorch and trained on an NVIDIA RTX 3090 GPU with 20GB of VRAM. The model is optimized using the Adam optimizer with an initial learning rate of 1×10^{-4} and a batch size of 4. The training process lasts for 80 epochs to ensure the stable convergence of the three-step unrolled reconstruction trajectory.

2.4. Configuration of RLD

In our comparative study, the physics-based RLD baseline utilizes a perfectly matched forward projection model (i.e., the same PSF matrix). We fix the RLD process to 8 iterations. Experimental results indicate that while RLD achieves higher resolution compared to earlier ray-optics-based benchmarks, a significant performance gap remains when compared to deep learning methods. This is attributed to the inherent trade-off between perspective and spatial resolution in LFM, as well as the extremely complex iterative evolution required by the physical model.

2.5. ICD Module and Distributional Adaptation Strategy

In our framework, the Mahalanobis distance (D_M) is employed for distribution estimation and OOD detection. Unlike the standard Euclidean distance, which assumes independence between feature dimensions, the Mahalanobis distance incorporates the covariance matrix Σ to account for correlations between biological feature statistics. This

provides a more robust metric for identifying subtle distribution shifts in complex microscopy data. When a sample is flagged as OOD, the ICD module signifies a potential decrease in reconstruction reliability and alerts the user to possible artifacts caused by the distribution shift. Furthermore, these flagged instances serve as priority candidates for data expansion. If the corresponding 3D ground truth can be subsequently acquired (e.g., via high-resolution wide-field cameras), these samples are utilized for supervised fine-tuning. This mechanism enables the TCD framework to adaptively expand its generalization boundary and maintain high-fidelity performance under evolving experimental conditions.

3. Additional Experimental and Visualizations

3.1. Visualization of DDPM

As shown in Figure 1, we present the 3D reconstruction results of DDPM over $T = 500$ inference steps. Starting from random Gaussian noise, the model progressively restores volumetric structures, with intermediate results displayed every 50 steps. Although DDPM eventually produces reasonable 3D reconstructions, its long inference chain results in high computational cost and slow speed, motivating our efficient Three-step Conditional Diffusion (TCD) framework.

3.2. Visual Comparison with SOTA

As shown in Figure 2, we provide visual comparisons between TCD and comparison methods on the Tubulin and Bcell datasets. TCD better recovers volumetric details and edge structures while maintaining spatial consistency. In contrast, baseline methods often suffer from oversmoothing or structural distortions in dense or detail-rich regions, whereas TCD consistently produces clearer and more coherent 3D structures.

3.3. Comparison of Noise-Level Robustness

As shown in Table 1, we compare the performance of different methods under various noise levels. To generate these levels, we add additive Gaussian noise to each clean image LF. The required noise variance σ_N^2 is dynamically computed for each image based on its own signal variance $\mathrm{Var}(LF)$ to achieve a specific target Signal-to-Noise Ratio (SNR_{dB}), defined as:

$$SNR_{dB} = 10 \cdot \log_{10}\left(\frac{\mathrm{Var}(LF)}{\sigma_N^2}\right) \tag{1}$$

TCD consistently outperforms baselines under medium-to-high SNR conditions, demonstrating strong noise robustness and structural preservation. Under extremely low SNR conditions (-10-0 dB), all methods degrade substantially, yet the overall trend indicates that the conditional diffusion design with noise-aware modeling improves stability under challenging noise conditions.

Table 1. **Quantitative comparison at different noise levels.**

Noise Level	Ours (TCD)		VCDNet		LFMNet	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
Clean	**41.12**	**0.900**	35.89	0.582	38.28	0.867
20dB	**40.90**	**0.897**	35.85	0.582	37.67	0.850
15dB	**40.66**	**0.894**	35.89	0.581	36.98	0.823
10dB	**40.17**	**0.880**	35.81	0.581	35.00	0.769
5dB	**38.44**	**0.842**	35.45	0.579	31.99	0.697
0dB	**34.57**	**0.762**	34.25	0.574	30.25	0.678
-5dB	30.89	0.682	**31.50**	0.565	30.42	**0.720**
-10dB	26.84	0.588	28.18	0.549	**32.18**	**0.721**

3.4. Visualization on the Mixed Dataset

As shown in Figure 4, We present 3D reconstruction results obtained by training on the mix biological dataset. TCD produces sharper details, more coherent volumetric textures, and more reliable depth reconstruction across diverse sample types. Note that (1) RLD is based on a physical model and is not applicable to mixed-sample training, and (2) CWFA requires strict consistency in volume depth and spatial size, making it unsuitable for this testing scenario.

Table 2. **Quantitative comparison before and after fine-tuning.**

Method	**TCD**		**TCD Fine-Tune**	
Scene	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
Bcell	38.24	0.880	40.42	0.899
Dendrite	41.30	0.901	42.56	0.908
Mito	35.04	0.827	36.47	0.870
Tubulin	29.94	0.791	44.65	0.981

3.5. ICD-based Fine-tuning Results

As shown in Table 2, for the OOD samples detected by ICD, PSNR and SSIM are both improved after fine-tuning, with particularly significant gains on structurally complex samples, indicating enhanced adaptability of the model to structural variations.

As shown in Figure 3, ICD fine-tuning reduces the average loss to 10^{-3} in nearly half the iterations compared to training from scratch, reflecting a two-fold acceleration in convergence and improved training efficiency.

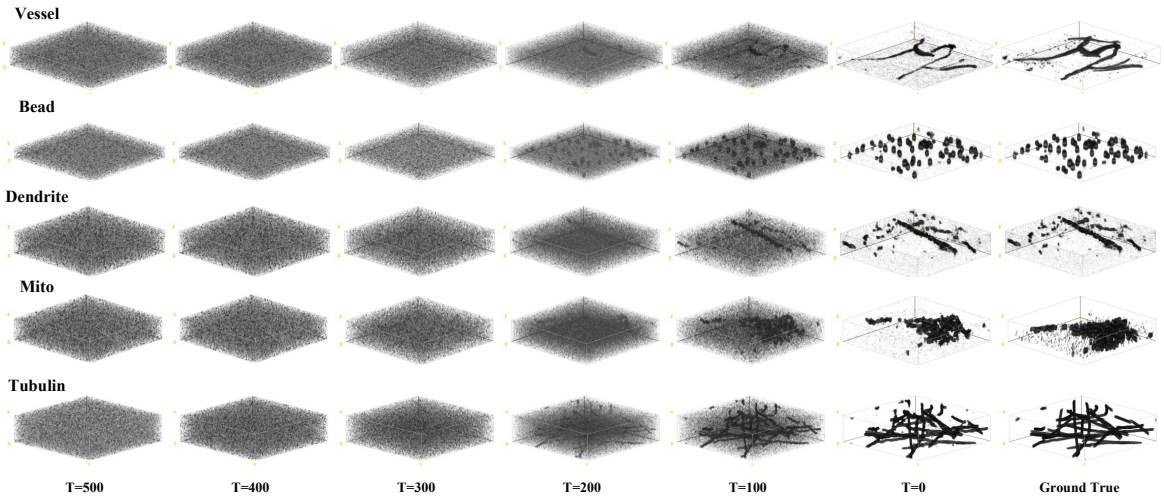

Figure 1. **Inference results of the DDPM-based model with $T = 500$ steps.**

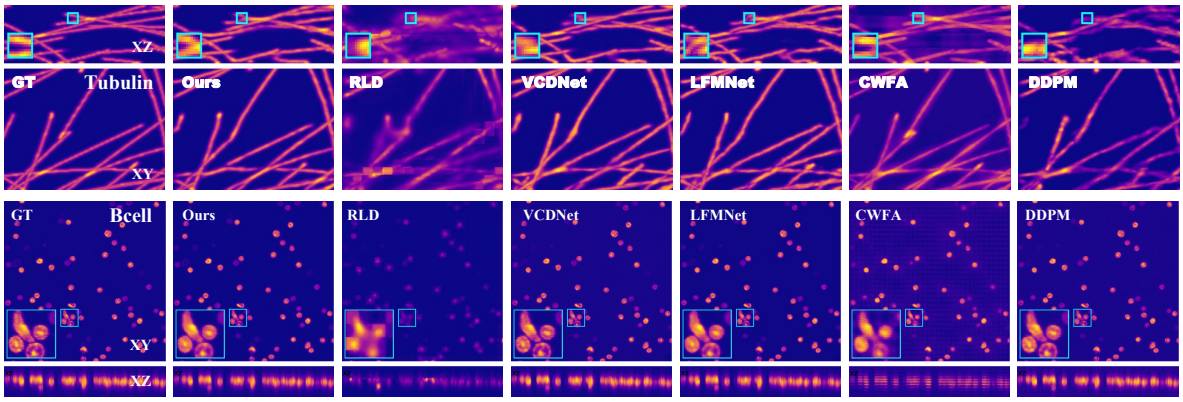

Figure 2. **Qualitative comparisons on the Tubulin, and Bcell datasets.**

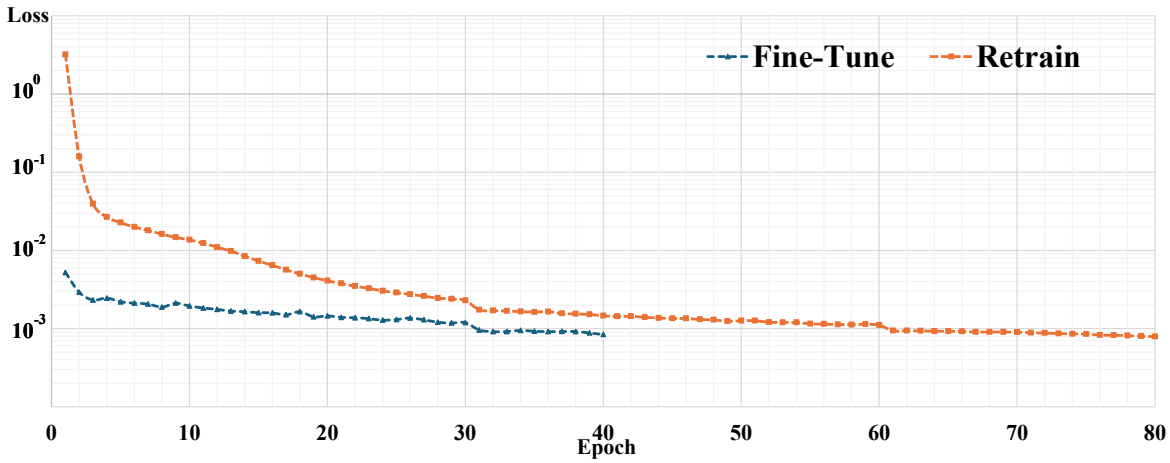

Figure 3. **Fine-tuning operation and loss convergence curve.**

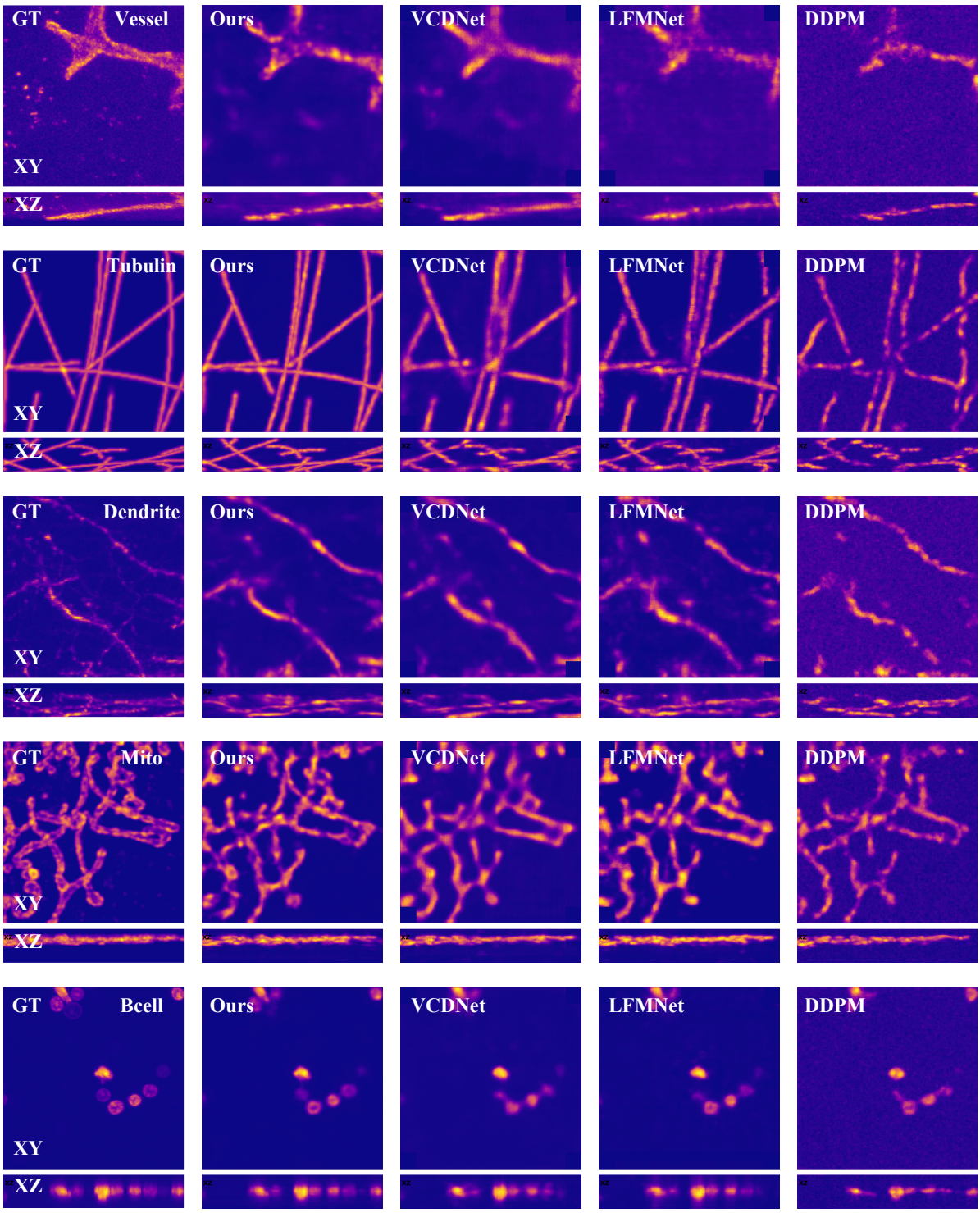

Figure 4. **Qualitative comparisons obtained by different methods in mixed datasets.**