V2V3D: View-to-View Denoised 3D Reconstruction for Light-Field Microscopy

Supplementary Material

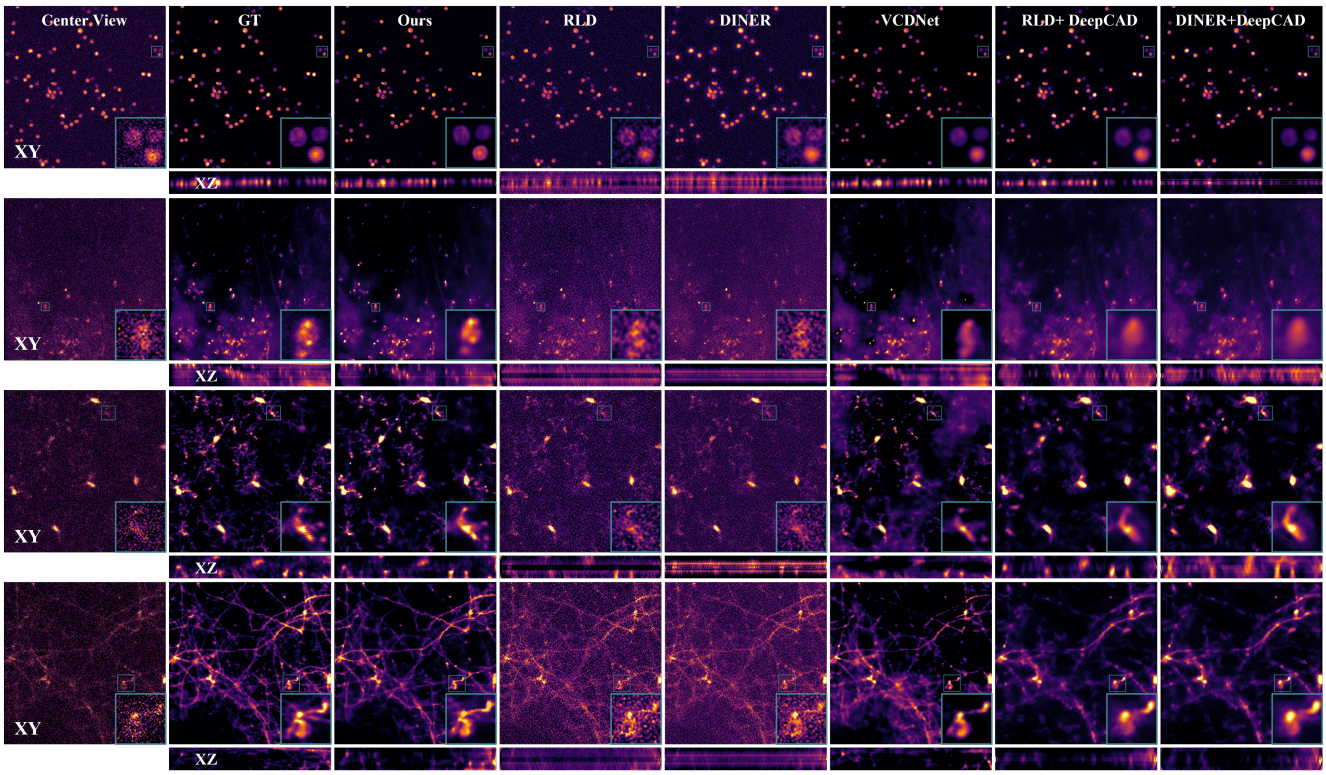

Figure 1. Qualitative comparisons on four biological samples, arranged from top to bottom, are as follows: B cells, vessels, microglia, and dendrites.

1. More Visual Comparison

We show more visual results in Figure 1. We compare the proposed method with state-of-the-art methods: RLD, DINER, VCDNet, RLD+DeepCAD, and DINER+DeepCAD. Our method consistently outperforms these comparison methods. Compared to unsupervised methods like RLD and DINER, our solution offers robust denoising capabilities and accurately reconstructs high-quality 3D volumes. When compared to supervised methods such as VCDNet, our approach shows enhanced generalization performance across diverse biological samples. Although denoised LFIs can boost the performance of RLD and DINER, they may also introduce additional artifacts and blurring in the reconstructed 3D volumes.

2. Impact of View Numbers in the Subsets

In this study, we propose a view-to-view denoised LFM 3D reconstruction framework, enabling the generation of high-quality 3D signals without requiring ground truth data. In the original V2V3D, we divide all views into two subsets of equal size. Here, we conduct experiments to examine the impact of the number of views in the subset. Specifically, we have tried the following divisions: [1,12], [2, 11], [3, 10], [4, 9], [5, 8], [6, 7]. Test results are presented in Figure 2. As the number of views in the minimal subset increases, the reconstruction performance improves. This aligns with expectations, as a small number of views in the subset leads to a significant degradation in the reconstruction quality, ultimately affecting the overall reconstruction performance.

3. Visualization of the Ablation Study

We have conducted ablation studies to investigate the role of the proposed V2V framework, the FFT loss, the feature alignment module, and the de-crosstalk loss. Quantitative results are presented in Table 3 (main text). In this subsection, we provide visualization results of the ablation study in Figure 3. These visual results further confirm the contribution of each component. Specifically, the V2V framework equips the network with denoising capabilities, facilitating the reconstruction of high-quality 3D volumes. The

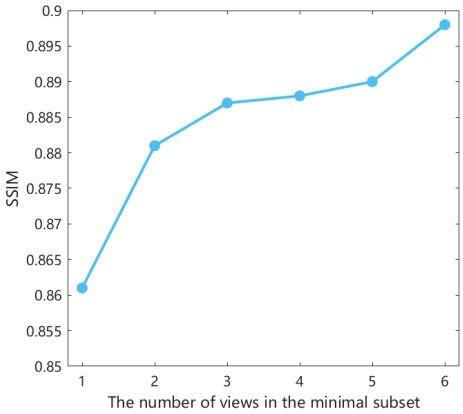

Figure 2. Impact of the number of views in the subset.

Table 1. Quantitative comparison of the performance and running time of different fusion strategies.

	Average	Max-pooling	Learnable aggregation
PSNR	39.05	38.86	39.11
Runtime(s)	0.413	0.417	0.483

feature alignment module enhances the reconstruction of fine details, highlighting its effectiveness. Meanwhile, the FFT loss and de-crosstalk loss aid the network in learning high-frequency components and suppressing artifacts, respectively. Additionally, to demonstrate the advantages of using a convolution kernel (with a diameter of 1) derived from the PSF for feature alignment, we attempted to directly use the flipped PSF instead. The results reveal a noticeable degree of detail blurring. This occurs because the PSF is essentially equivalent to a blur kernel.

4. Comparison of Fusion Strategies

We have explored more efficient fusion strategies. As shown in Tab.1, current methods (e.g., max-pooling, learnable aggregation) showed no significant improvement, so we chose the method with the lowest computational cost. Notably, unlike supervised methods with GT-guided aggregation, unsupervised fusion remains challenging.

5. Network Details

Figure 4 and Figure 5 illustrate the structure details of the encoder and decoder in our V2V3D framework. 'Conv3(1, 4)' denotes a 3 × 3 convolutional layer with 1 input channel and 4 output channels. 'Conv3(256, 128, 2)' denotes a 3 × 3 convolutional layer with 256 input channels, 128 output channels, and stride=2. 'AvgPool (2, 2)' refers to the Average pooling layer with kernel-size=2 and stride=2. 'MaxPool (2, 2)' denotes to the Max pooling layer with kernel-size=2 and stride=2. U and Z represent the number of input views and slices of the reconstructed volume, respectively.

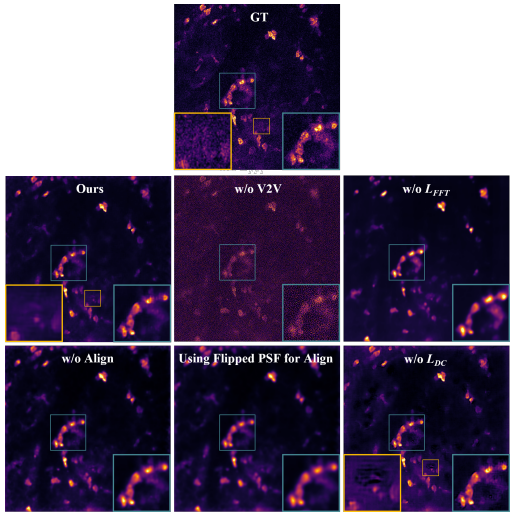

Figure 3. Visualization of the ablation study, presenting a slice of the reconstructed volume of neutrophils.

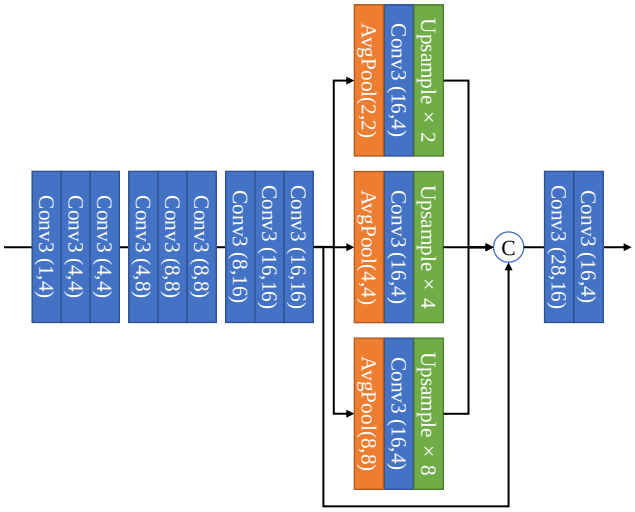

Figure 4. Network structure of the encoder.

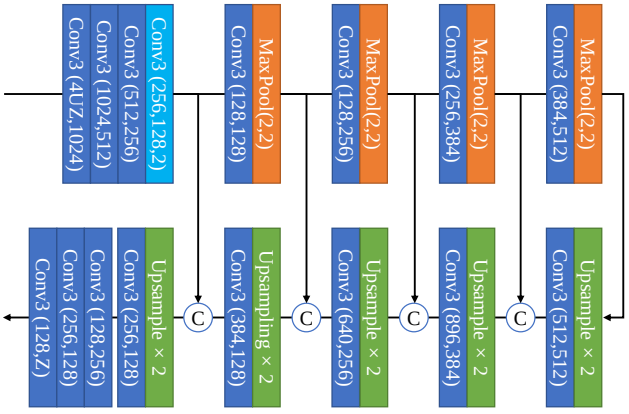

Figure 5. Network structure of the decoder.

Note that the activation functions (the LeakyReLU) are not shown in these figures. Specifically, our encoder consists

of multiple convolutional layers for deep feature extraction, and a pyramid structure is employed to capture multi-scale features, which are then concatenated to generate the final encoder features. Our decoder is built on a U-Net architecture, where the features are downsampled and upsampled five times.