

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

Jiayin Zhao^{1,2*}, Zhenqi Fu^{1*}, Tao Yu^{1†}, Hui Qiao^{1,2†}

¹Tsinghua University, Beijing 100084, China ²Shanghai AI Laboratory, China

zhao-jy23@mails.tsinghua.edu.cn, {fuzhenqi, ytrock, qiaohui}@mail.tsinghua.edu.cn

Abstract

Light field microscopy (LFM) has gained significant attention due to its ability to capture snapshot-based, large-scale 3D fluorescence images. However, existing LFM reconstruction algorithms are highly sensitive to sensor noise or require hard-to-get ground-truth annotated data for training. To address these challenges, this paper introduces V2V3D, an unsupervised view2view-based framework that establishes a new paradigm for joint optimization of image denoising and 3D reconstruction in a unified architecture. We assume that the LF images are derived from a consistent 3D signal, with the noise in each view being independent. This enables V2V3D to incorporate the principle of noise2noise for effective denoising. To enhance the recovery of high-frequency details, we propose a novel wave-optics-based feature alignment technique, which transforms the point spread function, used for forward propagation in wave optics, into convolution kernels specifically designed for feature alignment. Moreover, we introduce an LFM dataset containing LF images and their corresponding 3D intensity volumes. Extensive experiments demonstrate that our approach achieves high computational efficiency and outperforms the other state-of-the-art methods. These advancements position V2V3D as a promising solution for 3D imaging under challenging conditions. Our code and dataset will be publicly accessible at <https://joey1998hub.github.io/V2V3D/>.

1. Introduction

Light field microscopy (LFM) has emerged as a critical technology in a diverse array of biomedical applications, due to its unparalleled ability to capture high-resolution, three-dimensional microscopic scenes with exceptional precision and efficiency [35, 44]. The ability of simultaneously recording both spatial and angular information from the sample allows LFM to generate volumetric data, which

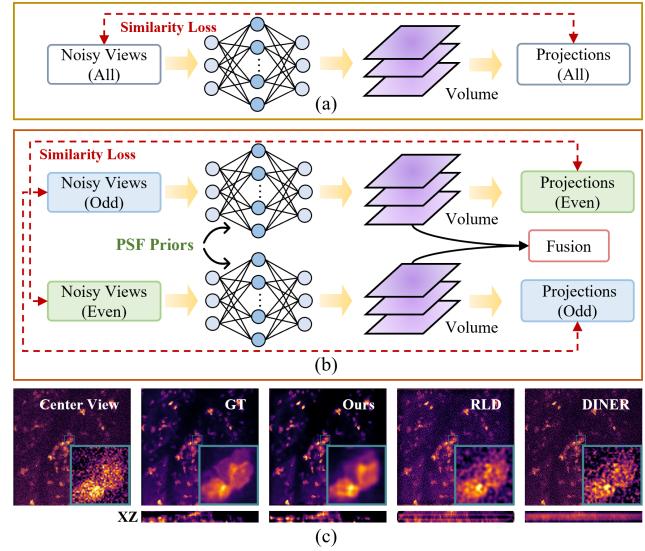


Figure 1. Background and concept of V2V3D. (a) Previous methods, such as VCDNet [42] and DINER [52], directly apply all views for reconstruction and lack physical priors in feature representation. Therefore, when reconstructing real-world noisy scenes, these methods usually generate results with conspicuous artifacts and blurriness. (b) The proposed method divides the noisy views into two non-overlapping subsets and employs two networks to generate the corresponding volumes. Additionally, we incorporate PSF priors for feature alignment, thereby enhancing feature aggregation across views. (c) Through the aforementioned custom designs, our method achieves state-of-the-art performance.

is particularly useful in dynamic biological processes where depth and temporal resolution are both crucial [7, 11, 47].

The most classical LFM reconstruction methods can generally be divided into two main categories: Richardson-Lucy deconvolution-based (RLD-based) approaches and learning-based solutions. Specifically, RLD-based methods [18, 29] rely on a computationally expensive and iterative recovery process, which severely limits the overall throughput of LFM reconstruction. This limitation makes them less suitable for long-duration or real-time applications. With the rapid progress of deep learning, learning-based algorithms [26, 42, 52], predominantly supervised learning methods, have emerged to enhance the speed and

*Co-first author

†Corresponding author

quality of LFM reconstruction. However, due to limited generalization capabilities, these algorithms are more suitable for scene-specific reconstruction.

Moreover, fluorescence microscopy of live cells requires gentle conditions, often necessitating low-light conditions that can result in substantial noise [37]. Existing LFM reconstruction algorithms directly utilize all available views to guide training. These methods implicitly average the noise across limited views, resulting in noticeable artifacts in the output. Using the pixel-wise independence of noise [19], noise-to-noise-based (N2N) methods can effectively reduce noise by learning a mapping between coordinate-matched image pairs. For light field images (LFIs) with significant noise, previous methods [44, 50] typically employ a N2N-based temporal denoising method [22] to obtain high signal-to-noise ratio (SNR) LFIs before reconstruction. However, this solution is suboptimal as it may lead to reduced temporal resolution and requires substantial temporal data, making it entirely unsuitable for snapshot applications. In fact, methods that generate paired noisy images from adjacent frames [22, 46] or adjacent pixels [2, 24] inevitably result in a substantial reduction in either temporal or spatial resolution [37].

In contrast to existing reconstruction methods that utilize pixel-to-pixel paired images for pre-denoising, we propose a view2view-based simultaneous denoising and 3D reconstruction framework, named V2V3D, which employs view-to-view paired noisy images (pixel-unmatched) as inputs and outputs. The key to achieving view-to-view denoising lies in establishing a mapping from one image space to another image space, which is physically consistent with the pipeline of unsupervised LFM 3D reconstruction [52]. As illustrated in Figure 1, the proposed V2V3D splits the views into two subsets, with each subset processed by a separate network to generate the corresponding volumes for fusion. Self-supervised losses are employed between the two branches to facilitate both reconstruction and denoising. Furthermore, V2V3D incorporates PSF priors for feature alignment, effectively warping coordinate-unmatched features into coordinate-matched features. This improvement enhances feature aggregation across different views, thereby further boosting the reconstruction performance in detail-rich areas. Extensive experiments demonstrate that V2V3D outperforms state-of-the-art methods in both noise removal and detail preservation, positioning it as a promising solution for robust snapshot 3D imaging across both microscopic and macro-scale scenarios. The main features of V2V3D are summarized as follows:

- A view2view-based simultaneous denoising and 3D reconstruction framework: V2V3D splits all views into two non-overlapping subsets and utilizes two separate networks to reconstruct the corresponding volumes. Using physical priors, it performs forward projection into sim-

ulated views, ensuring that the input views and the supervision views are different subsets. The network is optimized by minimizing the differences between the projected views and the real-captured views.

- A novel wave-optics-based feature alignment approach: We transform the PSF used in wave optics for forward propagation into convolution kernels for feature alignment, while also eliminating the blurring effects of the PSF. This feature alignment method enables efficient feature aggregation across different views, thereby supporting the recovery of high-frequency details.
- A light field dataset for quantitative evaluation: The ground-truth 3D intensity volumes, acquired via fluorescence microscopy, consists of 1618 high-resolution focal stacks. Then we utilize the principle of 2pSAM [50] to generate the corresponding LF images, as it has been validated to provide high-resolution imaging of deep tissues, particularly in terms of axial resolution.

2. Related Work

2.1. Traditional LFM 3D Reconstruction

Light field imaging was initially applied to macro scenarios [34]. Due to the ability of LF cameras to simultaneously capture spatial and angular information, they are commonly utilized in tasks such as 3D reconstruction [8, 33], image super-resolution [4, 6, 40], and depth estimation [9, 15, 21]. Its application in microscopy began in 2006 [20]. Then, Broxton *et al.* introduced wave optics to model the PSF of LFM [3] and applied the RLD method [5] to LFM. In 2019, Lu *et al.* proposed a RLD-based method [29] in phase space for LFM, effectively enhancing the reconstruction quality and convergence speed. Furthermore, Wu *et al.* proposed a unique scanning approach in LFM, simultaneously enhancing spatial and angular resolution with reduced phototoxicity [44]. However, conventional fluorescence microscopy struggles to achieve near-diffraction-limited resolution in deep tissue due to refractive index inhomogeneity and scattering [16, 53]. Two-photon microscopy (TPM) overcomes these issues through its longer wavelength and localized nonlinear excitation [10]. A recent innovation, 2pSAM [50], combines TPM with angular-scanning LF measurement to achieve near-diffraction-limited imaging with reduced photodamage.

However, these advanced LF imaging frameworks rely on RLD-based reconstruction algorithms, which limit the practicality of LFM. These algorithms iteratively correct the reconstructed volume based on Poisson assumption, essentially averaging the noise from all views in the output. This results in significant artifacts and image smoothing, while the iterative approaches are also computationally inefficient. Although more sophisticated RLD-based method has emerged for structured illumination microscopy [49], its

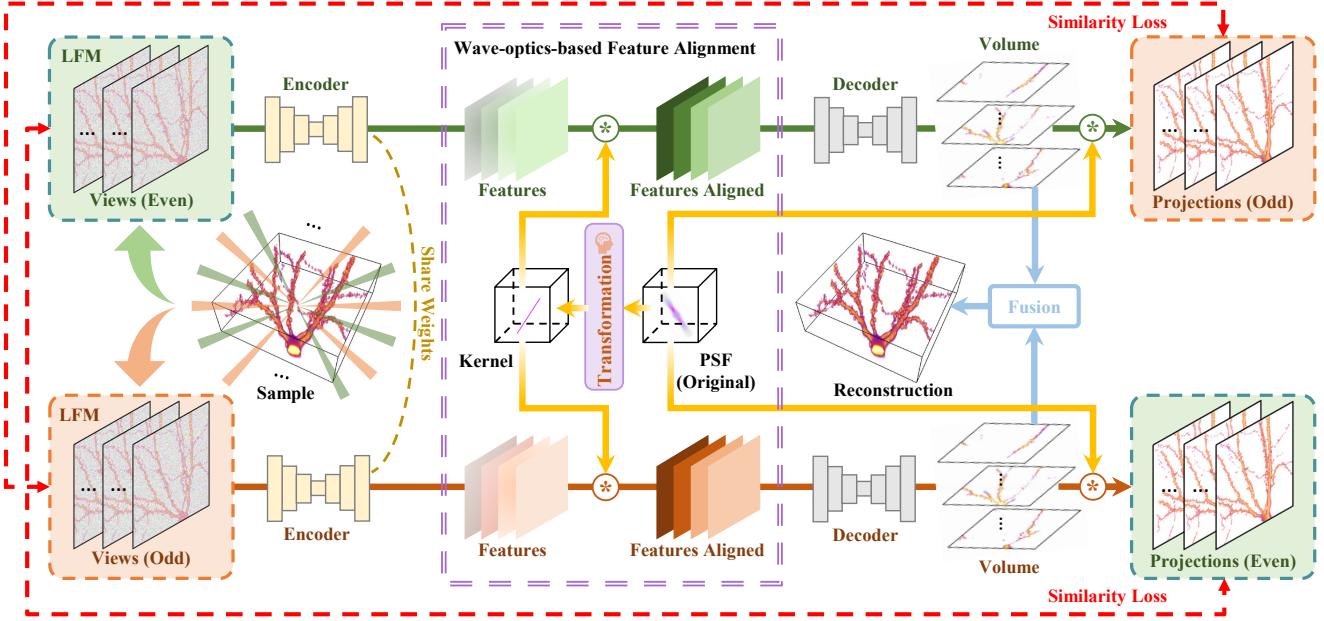


Figure 2. The overall framework of V2V3D, which divides all views into two subsets, with each subset generating a corresponding volume that collaborates to effectively reduce noise. \otimes denotes the 2D convolution operation. Additionally, V2V3D incorporates a novel wave-optics-based feature alignment technique, leveraging PSF priors to enhance the recovery of high-frequency information.

reliance on sparse priors and the use of strong regularization restrict their applicability to all microscopic contexts.

2.2. Deep-learning-based LFM 3D Reconstruction

Although recent deep-learning-based LFM reconstruction methods [26, 42] significantly improve the reconstruction efficiency compared with the RLD-based methods, their reconstruction quality is still far from practical due to poor generalization capacity and the lack of real-captured high-resolution training data. Additionally, convolutional networks such as VCDNet [42] concatenate all views as separate channels, where significant intensity variation occurs at the same location. Due to the limited receptive field of convolutional networks, this results in ineffective information aggregation and complicates the recovery of high-frequency details. Since 2020, implicit neural representation (INR) becomes a hot tool in the computer vision and graphics community for its superior performance on tasks like novel view synthesis [1, 32, 45], 3D reconstruction [14, 28, 51] and physical simulation [36, 38]. In recent developments, INR-based methods have emerged in the field of LFM. For example, DeCAF [27] has demonstrated the ability to eliminate the missing cone problem, it is too slow for use in long-term observation (e.g., DeCAF needs 20 hours to reconstruct single volume). DINER [52] significantly enhances reconstruction accuracy compared to DeCAF. However, its efficiency is still relatively lower when compared to convolutional networks, and it performs poorly in handling noise.

2.3. Denoising for Microscopy

Fluorescence microscopy of live cells necessitates gentle imaging conditions and sufficient spatiotemporal resolution, often resulting in a limited photon budget [13, 30, 37, 43]. To compensate for this constraint, improving the SNR is crucial for accurate LFM reconstruction. However, obtaining a sufficient collection of clean images for supervised learning poses significant challenges, particularly in live-cell applications. To remove noise without clean images, N2N-based methods [2, 12, 17, 19] learn mappings between pairs of independently degraded versions of the same image, achieving performance comparable to supervised methods. N2N-based methods have appeared in the field of TPM [22, 24]. For example, DeepCAD [23, 25] employs a self-supervised data generation process that assumes adjacent frames in a continuous imaging video share the same underlying content. However, in light field measurements, the considerable differences between adjacent views pose a challenge for N2N-based methods, which rely on the coordinate-matched image pairs. The variation in intensity at the same location can lead to noticeable artifacts.

3. Method

In this section, we first provide a brief introduction of LFM 3D imaging. Next, we detail the view-to-view framework and the feature alignment mechanism, emphasizing their crucial role in both denoising and reconstruction. Finally, we present the network architecture and loss functions.

3.1. Preliminaries

LFM employs 2D angular scanning techniques, such as LED multi-angle illumination and microlens arrays, to achieve high-speed 3D imaging with subcellular resolution. To simulate real-world LFM imaging, we developed a mathematical model that captures the entire process of light field imaging. First, we derive the point spread function (PSF) representation by modeling the light propagation process within a wave optics framework. This model encompasses the entire journey from the laser output to the objective plane. We define the direction of light propagation as the Z-axis and sample z points, with the intensity of each point represented as $I_{x,y,z}$. As illustrated in Figure 2, there are U beams illuminating the sample from different angles, each modeled as a PSF, denoted as $PSF_{u,x,y,z}$. Then the captured light field image (LFI) is represented as:

$$LFI_{u,x,y} = \sum_z (I_{x,y,z} * PSF_{u,x,y,z}), \quad (1)$$

where $*$ denotes the 2D convolution operation.

3.2. View-to-View-based LFM 3D Reconstruction

Fluorescence microscopy of live cells requires gentle conditions, often necessitating low-light environments that can result in substantial sensor noise. Due to the absence of noise modeling, current LFM reconstruction algorithms typically average noise across views in their results. However, this approach is ineffective at managing severe noise and may introduce significant artifacts. The N2N-based denoising methods leverage the inherent property of neural networks to avoid generating random noise. This characteristic facilitates effective noise reduction by mapping between pairs of coordinate-matched, noise-independent degraded versions of images, thereby preserving the underlying consistent signals. However, as illustrated in Eq. 1, the projection matrices (i.e., PSFs) that map the 3D information to 2D space differ for each view. As a result, the 2D coordinates of photons emitted from the same point in a 3D sample differ across different views, preventing the direct application of N2N-based methods for noise removal.

In this study, we propose a view2view-based framework that incorporates the principle of N2N [19] for denoising, enabling the reconstruction of high-quality 3D signals without ground truth data. We assume that the LFIs are fundamentally derived from a consistent 3D signal, with the noise in each view being independent. Our approach is to reconstruct a 3D signal using information from several views and then generate the remaining views using Eq. 1). This process generates pairs of coordinate-matched and noise-independent LFIs, thereby satisfying the N2N assumptions for effective denoising. Specifically, we divide all U views equally into two non-overlapping subsets, U_1 and U_2 . This partitioning strategy offers two key advantages: 1) All

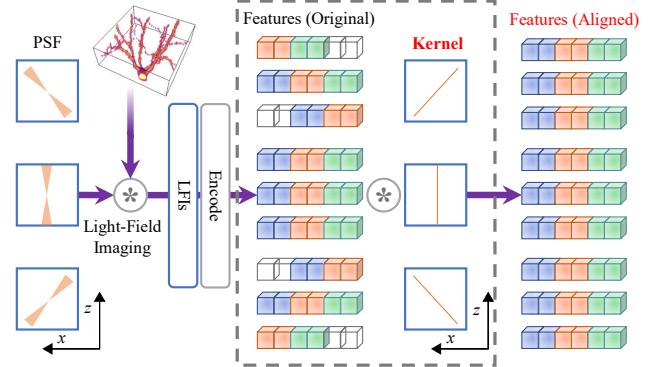


Figure 3. The diagram of the proposed wave-optics-based feature alignment module. The features extracted from different views are misaligned in the spatial dimension. To address this, we use kernels generated from the PSFs to align these features, thereby facilitating subsequent feature aggregation.

views are engaged per iteration, with one subset as input and the other for supervision; 2) The disjoint input/output pairing prevents trivial identity mappings, compelling the network to generate noise-free views by optimizing towards the statistical expectation of the target distribution.

As illustrated in Figure 2, the simulated view u_2 in subset U_2 can be generated using the subset U_1 , as expressed by

$$\hat{LFI}_{u_2,x,y} = \sum_z (f(LFI_{U_1,x,y}) * PSF_{u_2,x,y,z}), \quad (2)$$

where $f(\cdot)$ indicates a U-Net for reconstructing $\hat{I}_{x,y,z}$. Similarly, the simulated view u_1 in subset U_1 can be generated using the subset U_2 , as expressed by

$$\hat{LFI}_{u_1,x,y} = \sum_z (f(LFI_{U_2,x,y}) * PSF_{u_1,x,y,z}). \quad (3)$$

This network is optimized by minimizing the difference between the real-captured and simulated LFIs, yielding a model capable of high-quality reconstruction and denoising. Moreover, we implement two branches for reconstruction and merge the 3D signals produced by both branches to obtain the final reconstruction result.

3.3. Wave-optics-based Feature Alignment

The primary reason for the insufficient reconstruction quality of convolutional networks such as VCDNet is that they concatenate all views as separate channels and directly input them into the network. Due to the limited receptive field of convolutional networks, significant intensity variations across different channels at the same location can hinder effective information aggregation and impede the enhancement of reconstruction performance in detail-rich areas.

In RLD-based methods [18, 29], the information from the error map calculated between the simulated and real-captured LFIs can be progressively updated into the reconstructed 3D signal through back-projection, as expressed by

$$\Delta I_{update} = Error(LFI, \hat{LFI}) * PSF^{-1}, \quad (4)$$

Table 1. Quantitative comparison with state-of-the-art methods on the synthetic dataset. The best results are highlighted in bold.

Scene	RLD		VCDNet		DINER		DeepCAD+RLD		DeepCAD+DINER		Ours	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
B cells	27.13	0.507	45.15	0.972	28.01	0.481	29.05	0.896	32.58	0.859	38.73	0.981
Dendrites 1	27.30	0.516	30.41	0.709	26.85	0.386	30.16	0.734	30.47	0.806	30.92	0.768
Dendrites 2	34.76	0.819	36.25	0.851	33.14	0.740	34.88	0.866	34.82	0.857	36.65	0.897
Neutrophils	34.28	0.783	30.85	0.555	26.59	0.345	36.56	0.787	36.38	0.928	39.94	0.851
Microglia 1	30.43	0.800	40.16	0.960	26.52	0.560	32.15	0.895	35.53	0.958	38.35	0.966
Microglia 2	30.41	0.789	40.84	0.954	28.42	0.644	32.44	0.902	35.36	0.940	39.19	0.964
Neurons 1	31.52	0.697	37.44	0.747	31.12	0.625	33.19	0.821	33.17	0.825	36.39	0.783
Neurons 2	29.60	0.586	32.74	0.631	28.69	0.504	31.57	0.761	30.75	0.707	32.92	0.806
Neurons 3	37.39	0.839	42.85	0.870	36.70	0.795	40.70	0.920	39.72	0.921	43.30	0.932
Vessels 1	34.86	0.758	42.02	0.817	30.40	0.482	42.18	0.926	42.22	0.952	44.29	0.952
Vessels 2	36.46	0.838	36.06	0.645	29.23	0.459	46.63	0.964	40.96	0.977	48.91	0.976
Average	32.19	0.721	37.71	0.792	29.60	0.547	35.41	0.861	35.63	0.885	39.05	0.898

where PSF^{-1} can be obtained by flipping the PSF in two-dimensional space, as expressed by

$$PSF_{u,x,y,z}^{-1} = PSF_{u,-x,-y,z}. \quad (5)$$

Inspired by the back-projection technique used in RLD, we propose a wave-optics-based feature alignment method that enhances effective feature aggregation across different views. After extracting features for each view, we apply back-projection to warp all feature maps from their respective 2D spaces into a unified 3D space, facilitating improved feature aggregation. One might consider directly using PSF^{-1} for the back-projection of the feature maps. However, since the PSF acts as a low-pass filter in the frequency domain, this approach inevitably leads to feature blurring. To mitigate blurring effects, the PSF is converted into a convolution kernel with diameter 1 and weight 1, as illustrated in Figure 3. Specifically, we compute centroid coordinates of each PSF slice, set their values to 1, and set non-centroid positions to 0. Then the entire feature alignment process can be expressed as

$$Feature_{align} = Feature * Kernel_{PSF^{-1}}. \quad (6)$$

3.4. Network Architecture

As shown in Figure 2, the input of our method is the real-captured LFIIs, and the output is a high-resolution 3D intensity volume. The overall V2V3D reconstruction framework comprises two branches. Each branch includes: I) An encoder with a pyramid structure for feature extraction, with weights shared between two branches; II) A wave-optics-based feature alignment module that utilizes back-projection to warp all feature maps from different 2D spaces into a unified 3D space, facilitating feature aggregation; III) A U-Net-based decoder for generating a 3D volume from the aligned feature maps; IV) A forward projection module based on the physical modeling of the LFM system to

produce simulated LFIIs. The final reconstruction result is obtained by averaging the two volumes generated by the branches. Further details of the network architecture can be found in the supplementary materials.

3.5. Loss Functions

Although the MSE loss performs well in most scenarios, its effectiveness diminishes significantly in LFM reconstruction due to optical defocus, resulting in oversmoothing of high-frequency details. Therefore, we adopt the FFT Loss [48] to better recover high-frequency details. By utilizing the fast Fourier transform to map images from the spatial to the frequency domain, the FFT loss effectively balances the optimization of information across various frequencies. We define LFI_i as the value of real-captured projection pixel i , with the corresponding estimated pixel value denoted as \hat{LFI}_i . The MSE loss is defined as

$$L_{MSE} = \frac{\sum_i (LFI_i - \hat{LFI}_i)^2}{N}, \quad (7)$$

while the FFT loss can be expressed as

$$L_{FFT} = \frac{\left\| FFT(LFI) - FFT(\hat{LFI}) \right\|_2^2}{N}, \quad (8)$$

where $FFT(\cdot)$ indicates the fast Fourier transform and N is the total number of the pixels.

We also designed a regularization loss to mitigate artifacts caused by signal crosstalk along the Z-axis. In the reconstruction results of LFIIs with high brightness, a slice may exhibit excessively high intensity, while other slices could appear too dark, potentially falling below the sensor's background noise level. This discrepancy can lead to the emergence of significant artifacts. Therefore, we apply the de-crosstalk loss to penalize values that fall below the

Table 2. Efficiency comparison of V2V3D with other methods.

Method	PSNR	SSIM	Runtime (s)	Params (M)
VCDNet	37.71	0.792	0.047	87.98
RLD	32.19	0.721	7.34	-
DINER	29.60	0.547	61.4	62.92
V2V3D	39.05	0.898	0.413	210.77

background noise level of the sensor. Specifically, the de-crosstalk loss is defined as

$$L_{DC} = \sum_{x,y,z} \text{ReLU}(BG - \hat{I}_{x,y,z}), \quad (9)$$

where BG is the intensity of background noise, which can be estimated based on histogram analysis of all LFIs.

The final loss function is composed of three components: the MSE loss L_{MSE} , the FFT loss L_{FFT} , and the de-crosstalk loss L_{DC} . Thus, the overall training loss L_{all} is expressed as

$$L_{all} = L_{MSE} + \alpha L_{FFT} + \beta L_{DC}, \quad (10)$$

where α and β are weights. Empirically, we set $\alpha = 0.1$ and $\beta = 1$.

4. Experiments

In this section, we first introduce the experimental setup and datasets for evaluation. Then, we present both quantitative and qualitative comparisons with other SOTA methods. Finally, we conduct ablation studies to examine the various components of V2V3D.

4.1. Experimental Setup and Datasets

We confirmed the superior performance of V2V3D using both synthetic and real-world data. We utilized the principle of 2pSAM to obtain LFIs, as illustrated in Figure 2. This system features a distinctive “needle” beam for advanced light field imaging, facilitating both 2D spatial and angular scanning. This allows for high-speed, large-field 3D imaging at subcellular resolution. By rotating the mechanism, we can capture 13 LFIs of the 3D sample. For the synthetic dataset, we observed six types of biological scenarios using fluorescence microscopy, including B cells, dendrites, microglia, neurons, neutrophils, and blood vessels. Applying cropping and resizing operations, we obtained 1618 high-SNR 3D intensity volumes, each with a resolution of $512 \times 512 \times 39$. Then we used the generated PSFs of 2pSAM¹ and intensity volumes to perform physics-based forward projection, resulting in 1618 simulated light field images, each with a resolution of $512 \times 512 \times 13$. Finally, we introduced substantial Gaussian noise into the LFIs to simulate the actual imaging process. We selected 11 typical cases from the dataset to constitute the test set. For the

real-world dataset, we used the 2pSAM system to obtain LFIs. Note that we selected thick samples and shortened the exposure time to obtain low-SNR LFIs. We observed one static sample (brain-slice of mouse) and one live sample (neutrophils). Both of them consist of 100 frames, with each frame having a resolution of $512 \times 512 \times 13$. For the hardware configuration, we utilized an NVIDIA A100 GPU to handle the large-scale data and complex calculations involved in our study. For each method, we used PSNR and SSIM [41] to evaluate the accuracy of the reconstruction. Following the approach in [44], we subtract the background noise from both the reconstruction results and the ground truth before calculating the metrics.

4.2. Comparison on the Synthetic Dataset

We quantitatively compared our method with other SOTA ones on synthetic data, including an optimization-based method (RLD) [29], a supervised-learning-based method (VCDNet) [42] and a NeRF-based method (DINER) [52]. We retrained VCDNet on our dataset, using noisy LFIs as input and the noise-free 3D signal as supervision. Moreover, to validate the superiority of our framework in simultaneous reconstructing and denoising, we compared V2V3D with the unsupervised methods (RLD and DINER) that use pre-denoised LFIs as input. Note that we retrained DeepCAD [23] on our dataset for LFI denoising.

Table 1 reports the average metrics of all methods applied to noisy and pre-denoised synthetic data. Figure 4 and Figure 5 show the center views as well as the XY and XZ projections of the reconstructed and ground-truth 3D volumes. Table 2 offers a comprehensive quantitative comparison of all methods, including performance metrics and elapsed times. We can obtain the following conclusions: I) On the synthetic dataset, V2V3D demonstrates superior performance, particularly when compared to unsupervised methods. By leveraging the view-to-view reconstruction framework, our method is capable of recovering high-resolution, high-SNR 3D signals from LFIs with severe noise. Additionally, while using denoised LFIs as input can enhance the performance of RLD and DINER, it may also introduce more artifacts into the reconstructed 3D volumes; II) Due to the use of noise-free 3D signals as supervision, VCDNet can mitigate the impact of noise on reconstruction. However, by leveraging a physics-informed feature alignment module, V2V3D is able to reconstruct more high-frequency details than VCDNet; III) Benefiting from the convolutional framework, V2V3D shows superior reconstruction efficiency over optimization-based methods.

4.3. Comparison on Real-world Dataset

We conducted qualitative comparisons using two types of real data: one static sample (brain slice) and one live sample (neutrophils). For the static sample, we can obtain a high-

¹https://github.com/BBNCELi/2pSAM_recon

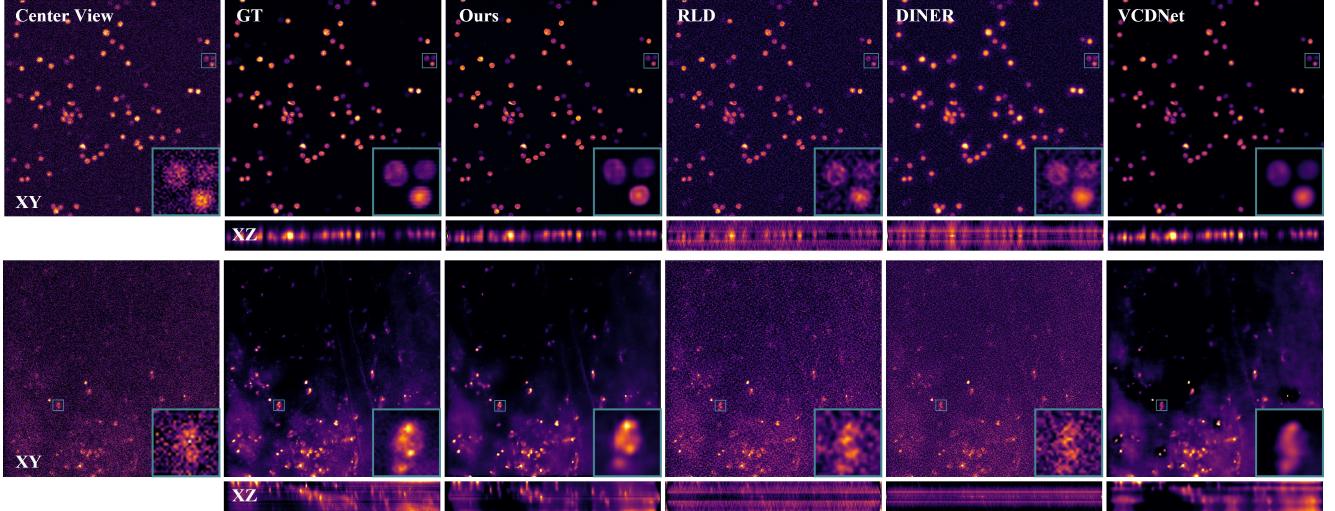


Figure 4. Qualitative comparisons on the synthetic dataset. Two biological samples arranged from top to bottom are B cells and vessels. Our solution delivers significantly higher quality, with less noise and sharper details.

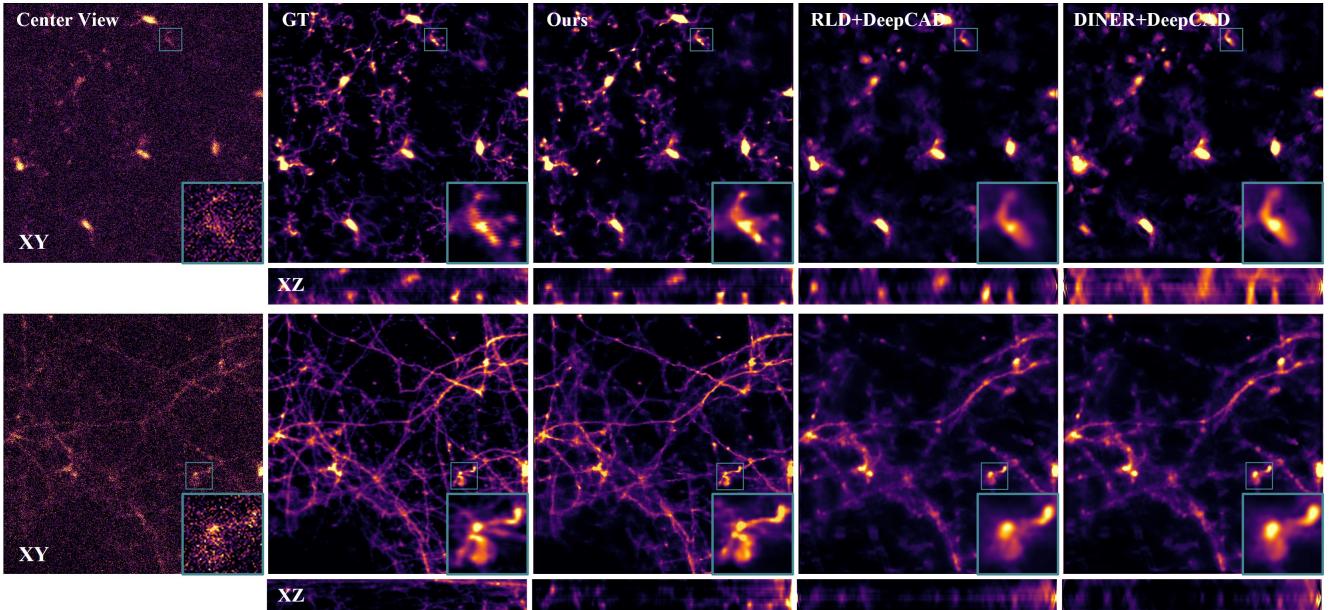


Figure 5. Qualitative comparisons on the synthetic dataset. Two biological samples arranged from top to bottom are microglia and dendrites. Our solution delivers significantly higher quality, with less noise and sharper details.

SNR reference center view through time averaging (100 frames), while for the live sample, we employed DeepCAD for denoising to acquire a relatively high-SNR reference center view. Figure 6 shows the center views, as well as the XY and XZ projections of the reconstructed 3D volumes. Severely affected by noise, both RLD and DINER exhibited significant deficiencies, as these methods essentially treat noise as valid signals in reconstruction. Consequently, the mean projection of the reconstructed volume closely resembles the center view. Since VCDNet was trained on a noisy dataset, it is capable of denoising to some extent. However, due to its poor generalization ability, the reconstruction results exhibit noticeable artifacts. Using pre-denoised

LFI_s from DeepCAD for reconstruction does improve the SNR of RLD and DINER. However, due to significant differences between adjacent views at the same position, directly applying the N2N-based denoising method can result in signal crosstalk and blurring in LFI_s, leading to the emergence of noticeable artifacts in final reconstruction results. Our method consistently outperformed other SOTA methods through several pivotal factors: the view2view framework effectively separates valid signals from severe noise, while the physics-informed feature alignment and FFT Loss boost the network's capacity to recover high-frequency details. These innovative techniques highlights V2V3D's potential applicability in real-world complex LFM imaging.

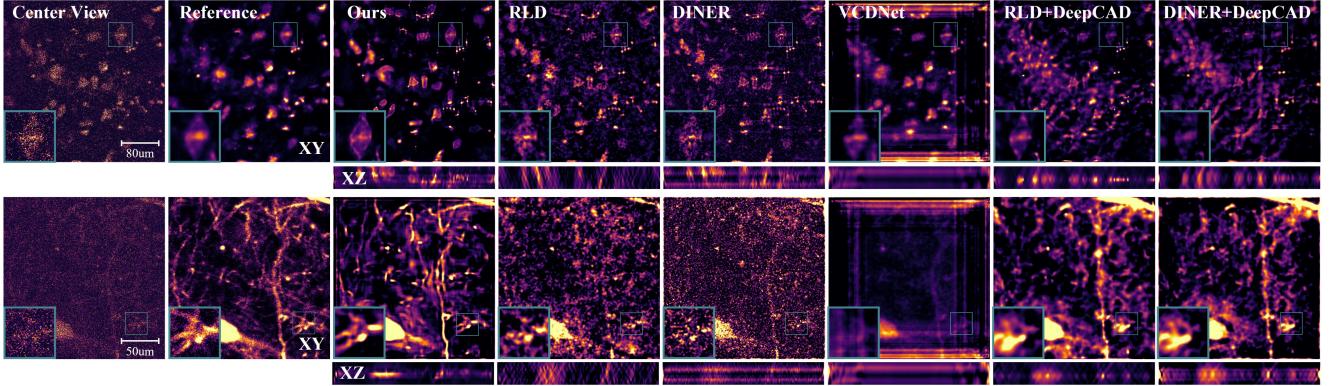


Figure 6. Qualitative comparisons on the real-world dataset. Two biological samples arranged from top to bottom are neutrophils and dendrites. For live sample (Neutrophils), we employ DeepCAD for denoising to acquire the reference center view. For static sample (Dendrites), we obtained a high SNR reference center view through time averaging.



Figure 7. Validation of denoising ability on macro-scale scenario.

4.4. Validation on Macro-scale Scenarios.

To validate the denoising ability of the view2view training strategy on macro-scale scenarios, we adapted V2V3D by building upon IBRNet [39]. Specifically, we sub-sampled the views of the *Fern* scene in the LLFF [31] by a factor of 4 and added severe Gaussian noise ($\mu = 0, \sigma = 50$). Then, we replaced the parts that generate views from one subset to another in both branches of V2V3D with IBRNet, and randomly sampled two non-overlapping subsets of views for training in each iteration. As shown in Figure 7, our method demonstrates robust performance on noisy macro images, outperforming both the pre-trained IBRNet and the one fine-tuned specifically for the *Fern* scene.

4.5. Ablation Study

Table 3 presents the contributions of every crucial component within V2V3D. Further visualizations of the ablation study can be found in the supplementary materials. Specifically, removing the view2view framework would cause the network to lose its denoising capability, resulting in a substantial drop in performance. Additionally, eliminating the feature alignment module results in a considerable loss of detail and overall image sharpness, further compromising the effectiveness of the reconstruction process. Furthermore, when the de-crosstalk loss is excluded, the artifacts caused by signal crosstalk in the reconstruction results increase significantly, degrading the quality of the output. Collectively, these findings underscore the critical importance of each component in enhancing the reconstruction quality, highlighting that each element plays a vital role in achieving optimal performance. We also explored other

Table 3. Ablation study on the V2V framework, feature alignment strategy, and the impact of FFT and de-crosstalk losses.

Metric	Ours	w/o V2V	w/o Align	w/o L_{FFT}	w/o L_{DC}
PSNR	39.05	30.81	38.25	37.23	36.09
SSIM	0.898	0.731	0.885	0.874	0.867

fusion strategies, e.g., max-pooling and learnable aggregation. Yet, as shown in the supplementary materials, these approaches brought no significant improvement.

5. Conclusions

Limitations and Future Works: Although V2V3D has achieved SOTA performance and shows significant promise in life science, there are still two directions for improvement: I) Develop more advanced fusion strategies rather than using straightforward averaging to further improve performance; II) Incorporate the optimization of the PSF during training to reduce the dependence of unsupervised methods on accurate imaging system models.

Conclusion: This study presents V2V3D, an view2view-based simultaneous denoising and 3D reconstruction framework for LFM. V2V3D divides all views into two non-overlapping subsets, each subset generating a corresponding volume and collaborating to remove noise. We also introduce a novel wave-optics-based feature alignment technique to improve reconstruction accuracy in detail-rich areas. Moreover, we introduce an LFM dataset to enable both quantitative and qualitative comparisons. We believe that V2V3D serves as a seminal exploration in simultaneous denoising and reconstruction, capable of stimulating more research within this burgeoning field.

6. Acknowledgement

This work was supported by National Key R&D Program of China (2023YFB3209700), National Natural Science Foundation of China (62322110, 62171255), and Postdoctoral Fellowship Program of CPSF (GZC20240836).

References

- [1] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5855–5864, 2021. 3
- [2] Joshua Batson and Loic Royer. Noise2self: Blind denoising by self-supervision. In *International Conference on Machine Learning*, pages 524–533. PMLR, 2019. 2, 3
- [3] Michael Broxton, Logan Grosenick, Samuel Yang, Noy Cohen, Aaron Andalman, Karl Deisseroth, and Marc Levoy. Wave optics theory and 3-d deconvolution for the light field microscope. *Optics express*, 21(21):25418–25439, 2013. 2
- [4] Zhen Cheng, Zhiwei Xiong, Chang Chen, Dong Liu, and Zheng-Jun Zha. Light field super-resolution with zero-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10010–10019, 2021. 2
- [5] DA Fish, AM Brinicombe, ER Pike, and JG Walker. Blind deconvolution by means of the richardson–lucy algorithm. *JOSA A*, 12(1):58–65, 1995. 2
- [6] Chen Gao, Youfang Lin, Song Chang, and Shuo Zhang. Spatial-angular multi-scale mechanism for light field spatial super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1961–1970, 2023. 2
- [7] Changliang Guo, Wenhao Liu, Xuanwen Hua, Haoyu Li, and Shu Jia. Fourier light-field microscopy. *Optics express*, 27(18):25573–25594, 2019. 1
- [8] Shuji Habuchi, Keita Takahashi, Chihiro Tsutake, Toshiaki Fujii, and Hajime Nagahara. Time-efficient light-field acquisition using coded aperture and events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24923–24933, 2024. 2
- [9] Kang Han, Wei Xiang, Eric Wang, and Tao Huang. A novel occlusion-aware vote cost for light field depth estimation. *IEEE transactions on pattern analysis and machine intelligence*, 44(11):8022–8035, 2021. 2
- [10] Fritjof Helmchen and Winfried Denk. Deep tissue two-photon microscopy. *Nature methods*, 2(12):932–940, 2005. 2
- [11] Xia Hua, Yujie Wang, Shuming Wang, Xiujuan Zou, You Zhou, Lin Li, Feng Yan, Xun Cao, Shumin Xiao, Din Ping Tsai, et al. Ultra-compact snapshot spectral light-field imaging. *Nature communications*, 13(1):2732, 2022. 1
- [12] Tao Huang, Songjiang Li, Xu Jia, Huchuan Lu, and Jianzhuang Liu. Neighbor2neighbor: Self-supervised denoising from single noisy images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14781–14790, 2021. 3
- [13] Xiaoshuai Huang, Junchao Fan, Liuju Li, Haosen Liu, Runlong Wu, Yi Wu, Lisi Wei, Heng Mao, Amit Lal, Peng Xi, et al. Fast, long-term, super-resolution imaging with hessian structured illumination microscopy. *Nature biotechnology*, 36(5):451–459, 2018. 3
- [14] Mude Hui, Zihao Wei, Hongru Zhu, Fei Xia, and Yuyin Zhou. Microdiffusion: Implicit representation-guided diffusion for 3d reconstruction from limited 2d microscopy projections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11460–11469, 2024. 3
- [15] Jing Jin, Junhui Hou, Jie Chen, Huanqiang Zeng, Sam Kwong, and Jingyi Yu. Deep coarse-to-fine dense light field reconstruction with flexible sampling and geometry-aware fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4):1819–1836, 2020. 2
- [16] Chui Kong, Yangzhen Wang, and Guihua Xiao. Neuron populations across layer 2–6 in the mouse visual cortex exhibit different coding abilities in the awake mice. *Frontiers in Cellular Neuroscience*, 17:1238777, 2023. 2
- [17] Alexander Krull, Tim-Oliver Buchholz, and Florian Jug. Noise2void-learning denoising from single noisy images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2129–2137, 2019. 3
- [18] Martin Laasmaa, Marko Vendelin, and Pearu Peterson. Application of regularized richardson-lucy algorithm for deconvolution of confocal microscopy images. *Biophysical Journal*, 100(3):139a, 2011. 1, 4
- [19] Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. Noise2noise: Learning image restoration without clean data. *International Conference on Machine Learning*, *International Conference on Machine Learning*, 2018. 2, 3, 4
- [20] Marc Levoy, Ren Ng, Andrew Adams, Matthew Footer, and Mark Horowitz. Light field microscopy. In *Acm siggraph 2006 papers*, pages 924–934. ACM SIGGRAPH, 2006. 2
- [21] Peng Li, Jiayin Zhao, Jingyao Wu, Chao Deng, Yuqi Han, Haoqian Wang, and Tao Yu. Opal: Occlusion pattern aware loss for unsupervised light field disparity estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2
- [22] Xinyang Li, Guoxun Zhang, Hui Qiao, Feng Bao, Yue Deng, Jiamin Wu, Yangfan He, Jingping Yun, Xing Lin, Hao Xie, et al. Unsupervised content-preserving transformation for optical microscopy. *Light: Science & Applications*, 10(1):44, 2021. 2, 3
- [23] Xinyang Li, Guoxun Zhang, Jiamin Wu, Yuanlong Zhang, Zhifeng Zhao, Xing Lin, Hui Qiao, Hao Xie, Haoqian Wang, Lu Fang, et al. Reinforcing neuron extraction and spike inference in calcium imaging using deep self-supervised denoising. *Nature methods*, 18(11):1395–1400, 2021. 3, 6
- [24] Xinyang Li, Xiaowan Hu, Xingye Chen, Jiaqi Fan, Zhifeng Zhao, Jiamin Wu, Haoqian Wang, and Qionghai Dai. Spatial redundancy transformer for self-supervised fluorescence image denoising. *Nature Computational Science*, 3(12):1067–1080, 2023. 2, 3
- [25] Xinyang Li, Yixin Li, Yiliang Zhou, Jiamin Wu, Zhifeng Zhao, Jiaqi Fan, Fei Deng, Zhaofa Wu, Guihua Xiao, Jing He, et al. Real-time denoising enables high-sensitivity fluorescence time-lapse imaging beyond the shot-noise limit. *Nature Biotechnology*, 41(2):282–292, 2023. 3

- [26] Yue Li, Yijun Su, Min Guo, Xiaofei Han, Jiamin Liu, Harshad D Vishwasrao, Xuesong Li, Ryan Christensen, Titas Sengupta, Mark W Moyle, et al. Incorporating the image formation process into deep learning improves network performance. *Nature Methods*, 19(11):1427–1437, 2022. 1, 3
- [27] Renhao Liu, Yu Sun, Jiabei Zhu, Lei Tian, and Ulugbek S Kamilov. Recovery of continuous 3d refractive index maps from discrete intensity-only measurements using neural fields. *Nature Machine Intelligence*, 4(9):781–791, 2022. 3
- [28] Zhen Liu, Hao Zhu, Qi Zhang, Jingde Fu, Weibing Deng, Zhan Ma, Yanwen Guo, and Xun Cao. Finer: Flexible spectral-bias tuning in implicit neural representation by variable-periodic activation functions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2713–2722, 2024. 3
- [29] Zhi Lu, Jiamin Wu, Hui Qiao, You Zhou, Tao Yan, Zijing Zhou, Xu Zhang, Jingtao Fan, and Qionghai Dai. Phase-space deconvolution for light field microscopy. *Optics express*, 27(13):18131–18145, 2019. 1, 2, 4, 6
- [30] Florian Luisier, Cédric Vonesch, Thierry Blu, and Michael Unser. Fast interscale wavelet denoising of poisson-corrupted images. *Signal processing*, 90(2):415–427, 2010. 3
- [31] Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (ToG)*, 38(4):1–14, 2019. 8
- [32] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 3
- [33] Ryoya Mizuno, Keita Takahashi, Michitaka Yoshida, Chihiro Tsutake, Toshiaki Fujii, and Hajime Nagahara. Acquiring a dynamic light field through a single-shot coded image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19830–19840, 2022. 2
- [34] Ren Ng, Marc Levoy, Mathieu Brédif, Gene Duval, Mark Horowitz, and Pat Hanrahan. *Light field photography with a hand-held plenoptic camera*. PhD thesis, Stanford university, 2005. 2
- [35] Nicolas C Pégard, Hsiou-Yuan Liu, Nick Antipa, Maximilian Gerlock, Hillel Adesnik, and Laura Waller. Compressive light-field microscopy for 3d neural activity recording. *Optica*, 3(5):517–524, 2016. 1
- [36] Yi-Ling Qiao, Alexander Gao, and Ming Lin. Neu-physics: Editable neural geometry and physics from monocular videos. *Advances in Neural Information Processing Systems*, 35:12841–12854, 2022. 3
- [37] Liying Qu, Shiqun Zhao, Yuanyuan Huang, Xianxin Ye, Kunhao Wang, Yuzhen Liu, Xianming Liu, Heng Mao, Guangwei Hu, Wei Chen, et al. Self-inspired learning for denoising live-cell super-resolution microscopy. *Nature Methods*, pages 1–14, 2024. 2, 3
- [38] Haoxiang Wang, Tao Yu, Tianwei Yang, Hui Qiao, and Qionghai Dai. Neural physical simulation with multi-resolution hash grid encoding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5410–5418, 2024. 3
- [39] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2021. 8
- [40] Yingqian Wang, Longguang Wang, Gaochang Wu, Jungang Yang, Wei An, Jingyi Yu, and Yulan Guo. Disentangling light fields for super-resolution and disparity estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):425–443, 2022. 2
- [41] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6
- [42] Zhaoqiang Wang, Lanxin Zhu, Hao Zhang, Guo Li, Chengqiang Yi, Yi Li, Yicong Yang, Yichen Ding, Mei Zhen, Shangbang Gao, et al. Real-time volumetric reconstruction of biological dynamics with light-field microscopy and deep learning. *Nature methods*, 18(5):551–556, 2021. 1, 3, 6
- [43] Martin Weigert, Uwe Schmidt, Tobias Boothe, Andreas Müller, Alexandr Dibrov, Akanksha Jain, Benjamin Wilhelm, Deborah Schmidt, Coleman Broaddus, Siân Culley, et al. Content-aware image restoration: pushing the limits of fluorescence microscopy. *Nature methods*, 15(12):1090–1097, 2018. 3
- [44] Jiamin Wu, Zhi Lu, Dong Jiang, Yuduo Guo, Hui Qiao, Yi Zhang, Tianyi Zhu, Yeyi Cai, Xu Zhang, Karl Zhanghao, et al. Iterative tomography with digital adaptive optics permits hour-long intravital observation of 3d subcellular dynamics at millisecond scale. *Cell*, 184(12):3318–3332, 2021. 1, 2, 6
- [45] Haiyang Ying, Baowei Jiang, Jinzhi Zhang, Di Xu, Tao Yu, Qionghai Dai, and Lu Fang. Parf: Primitive-aware radiance fusion for indoor scene novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17706–17716, 2023. 3
- [46] Guoxun Zhang, Xiaopeng Li, Yuanlong Zhang, Xiaofei Han, Xinyang Li, Jinqiang Yu, Boqi Liu, Jiamin Wu, Li Yu, and Qionghai Dai. Bio-friendly long-term subcellular dynamic recording by self-supervised image enhancement microscopy. *Nature Methods*, 20(12):1957–1970, 2023. 2
- [47] Yuanlong Zhang, Mingrui Wang, Qiyu Zhu, Yuduo Guo, Bo Liu, Jiamin Li, Xiao Yao, Chui Kong, Yi Zhang, Yuchao Huang, et al. Long-term mesoscale imaging of 3d intercellular dynamics across a mammalian organ. *Cell*, 2024. 1
- [48] Jiayin Zhao, Zhifeng Zhao, Jiamin Wu, Tao Yu, and Hui Qiao. Pnr: Physics-informed neural representation for high-resolution lfm reconstruction. *arXiv preprint arXiv:2409.18223*, 2024. 5
- [49] Weisong Zhao, Shiqun Zhao, Liuju Li, Xiaoshuai Huang, Shijia Xing, Yulin Zhang, Guohua Qiu, Zhenqian Han,

- Yingxu Shang, De-en Sun, et al. Sparse deconvolution improves the resolution of live-cell super-resolution fluorescence microscopy. *Nature biotechnology*, 40(4):606–617, 2022. 2
- [50] Zhifeng Zhao, Yiliang Zhou, Bo Liu, Jing He, Jiayin Zhao, Yeyi Cai, Jingtao Fan, Xinyang Li, Zilin Wang, Zhi Lu, et al. Two-photon synthetic aperture microscopy for minimally invasive fast 3d imaging of native subcellular behaviors in deep tissue. *Cell*, 186(11):2475–2491, 2023. 2
- [51] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):3170–3184, 2021. 3
- [52] Hao Zhu, Shaowen Xie, Zhen Liu, Fengyi Liu, Qi Zhang, You Zhou, Yi Lin, Zhan Ma, and Xun Cao. Disorder-invariant implicit neural representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1, 2, 3, 6
- [53] Warren R Zipfel, Rebecca M Williams, and Watt W Webb. Nonlinear magic: multiphoton microscopy in the biosciences. *Nature biotechnology*, 21(11):1369–1377, 2003. 2