XLFM-Former: Rapid 3D Reconstruction for Light Field Microscopy

Anonymous ICCV submission

Paper ID 8604

Abstract

Light field microscopy (LFM) has become an emerging tool in neuroscience for large-scale neural imaging in vivo, with XLFM (eXtended Light Field Microscopy) notable for its single-exposure volumetric imaging, broad field of view, and high temporal resolution. However, 3D reconstruction of neural activity from multi-view light field data is slow and computationally expensive, hindering real-time brain dynamics observation and broader applications. Deep learning could provide faster and more accurate solutions, but advancements are hampered by a lack of datasets, benchmarks, and effective architectures. We present eXtended Light Field Microscopy-Transformer (XLFM-Former), a transformer-based framework for real-time (> 30 volumes/second) and high-quality XLFM reconstruction. It uses a Swin Transformer encoder for hierarchical feature extraction and progressive upsampling for cross-level fusion. Key innovations include: (1) Masked View Modeling-Light Field (MVM-LF) pretraining for global view representation; (2) Physics-guided Point Spread Function (PSF) loss for optical consistency. When tested on the unique XLFM-zebrafish dataset, XLFM-Former outperforms current methods, achieving 7.7% higher PSNR while maintaining fine neural structures. To support future research, we release the benchmark dataset and evaluation metrics. Code and dataset available at: xxx.

1. Introducion

Light Field Microscopy (LFM) has emerged as a crucial technique for rapid volumetric imaging of nervous systems [19, 20, 36]. Notably, eXtended Light Field Microscopy (XLFM) [8], due to its graceful balance between speed, scale and resolution, is considered one of the most suitable LFM techniques for large-scale neural activity recording in several model organisms, including fish and mouse [2]. XLFM offers several advantages: 1)XLFM enables single-exposure acquisition of complete light field information at 100 Hz, whereas conventional microscopy techniques (e.g., two-photon microscopy [11], light-sheet microscopy [17]) require sequential layer-by-layer scanning, making it challenging to capture sub-second large-scale neural dynamics simultaneously. 2) The XLFM system incorporates a point spread function (PSF) that is approximately spatially invariant. Consequently, the reconstruction of volumes through 3D deconvolution is free from artifacts. 3) The rapid speed of XLFM allows real-time observation of large-scale population neural activity *in vivo*. Integrating volumetric imaging with optogenetic manipulation [4, 35] will enable optical brain-machine interface, namely closed-loop optical interrogation of brain-wide activity in both immobilized [28] and freely behaving animals [5, 8].

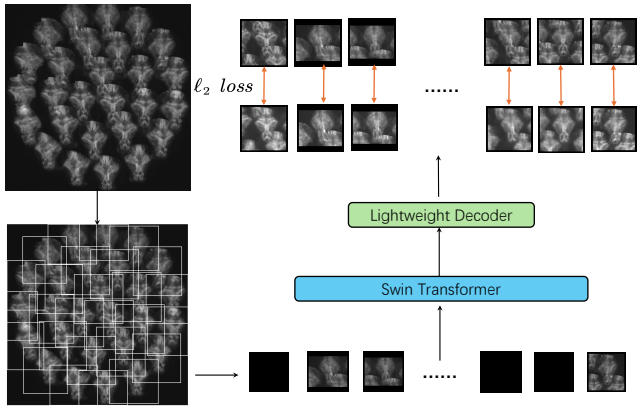

Figure 1. Our pretraining pipeline for XLFM. The raw light field acquired from the microscope is separated into 27 distinct viewpoints based on physical coordinates. With a 70% probability, we randomly mask a subset of these viewpoints and task the model with reconstructing them. The training is supervised by an ℓ_2 loss comparing the predicted and ground-truth views.

However, the distinctive optical sampling method employed by XLFM causes a speed bottleneck in 3D reconstruction, not data acquisition. Traditional approaches, such as iterative Richardson-Lucy deconvolution, are inherently slow, necessitating the development of more efficient reconstruction algorithms that can integrate both the physics priors and the expressiveness inherent in neural network models. Despite its great potential, existing deep learning

ICCV #8604

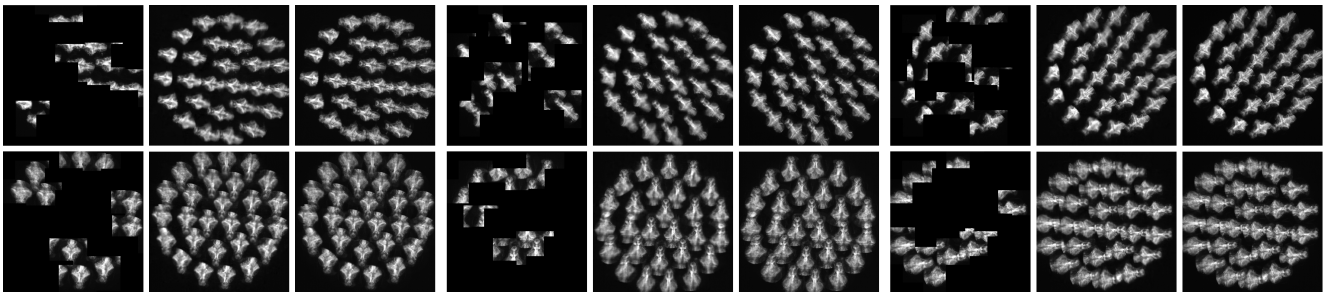

Figure 2. **Zebrafish multi-view light field images used for pretraining an encoder model.** For each triplet, we show the masked image (left), our MVM-LF regenerated image (middle), and the ground-truth (right).

approaches still face several challenges in XLFM reconstruction tasks. Firstly, high-quality annotation is costly. While XLFM sampling data is readily available, generating high-quality 3D reconstruction annotations (e.g., via Richardson-Lucy deconvolution [24]) is highly expensive, limiting the scalability of supervised learning methods. Secondly, multi-view geometric modeling remains insufficient. Although XLFM captures light field information using a microlens array, existing approaches struggle to fully exploit multi-view data for enhancing 3D reconstruction quality. Additionally, the lack of physical-optical priors poses another challenge. Most existing methods rely purely on data-driven approaches, neglecting the physical constraints of microscopic optical imaging, which may lead to reconstruction results that deviate from optical imaging principles.

To address these issues, we propose eXtended Light Field Microscopy-Transformer (XLFM-Former), an end-to-end Transformer framework specifically optimized for XLFM tasks. XLFM-Former employs hierarchical feature encoding and a window attention mechanism to enhance its modeling capability. The XLFM-Former encoder, constructed based on the window attention mechanism, extracts semantic features at four different resolutions and scales, ensuring effective modeling of both local and global information. Subsequently, a CNN-based decoder is employed for feature fusion to restore high-resolution 3D structural information. The decoder is followed by a reconstruction head, which directly computes the final 3D reconstruction results. To further enhance model performance, we introduce a self-supervised pretraining strategy for the Transformer encoder. Specifically, since XLFM acquires light field data from 27 different viewpoints through the microlens array, we propose a novel self-supervised learning task—Masked View Modeling-Light Field (MVM-LF)—to improve the model's understanding of multi-view light field information. During pre-training, 70% of the viewpoint data is randomly masked, and the model is provided with only 30% of the viewpoints, requiring it to predict the content of the masked viewpoints (See Figure 1 for an overview). This encourages the model to learn the underlying geometric relationships within the light field data. By leveraging the intrinsic structural priors of XLFM, this approach reduces reliance on costly annotations while significantly improving model generalization. To get a qualitative sense of our reconstruction task, see Figure 2 and 3.

Beyond data-driven learning, we introduce a PSF-based physical constraint loss to ensure that the reconstructed 3D structures conform to the optical system's physical principles. The Point Spread Function (PSF) [1], derived from the microscope's optical system, characterizes the diffraction and spreading behavior of a point source within the imaging system. Unlike conventional loss functions (e.g., L1/L2, SSIM), the PSF loss function forward-projects the 3D reconstruction results back into the original light field, ensuring that model predictions adhere to the physical imaging process. This enables XLFM-Former to effectively learn the optical characteristics of light field data, improving reconstruction quality while minimizing artifacts that violate optical principles. Experimental results demonstrate that XLFM-Former outperforms existing state-of-the-art frameworks, achieving superior reconstruction quality on XLFM tasks.

Our main contributions are summarized as follows:

- We establish a standardized framework for the XLFM reconstruction task and construct the first benchmark dataset for XLFM reconstruction. For the first time, we present a large-scale XLFM dataset comprising 22,581 images, including three free-swimming zebrafish, seven immobilized zebrafish, and six test zebrafish, captured at varying acquisition rates (10 fps / 1 fps). This dataset establishes a standardized and reproducible benchmark, facilitating the evaluation and advancement of XLFM reconstruction methodologies.
- We design the eXtended Light Field Microscopy-Transformer (XLFM-Former) framework, integrating Swin Transformer's hierarchical modeling capabilities to effectively capture both local and global information in XLFM data while adapting to light field viewpoint modeling tasks.

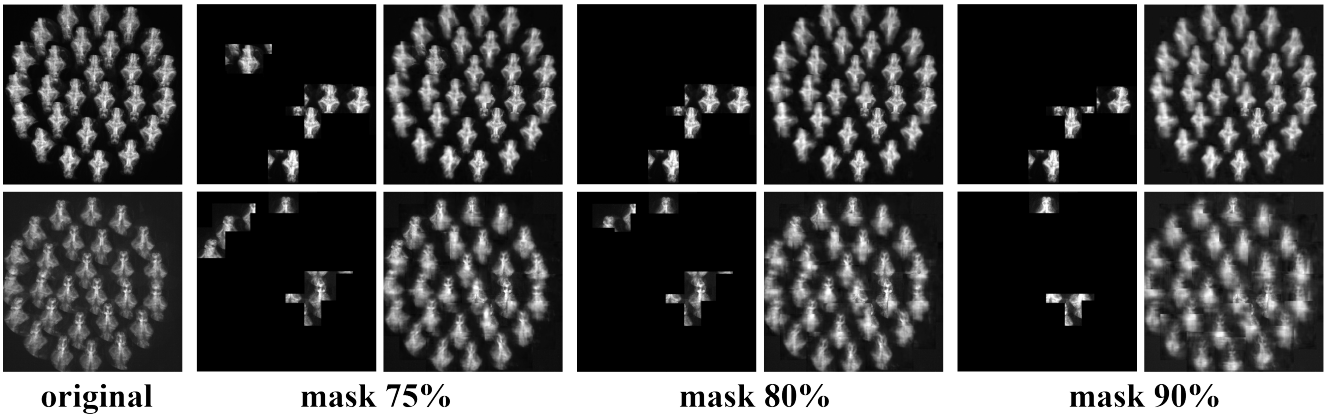

Figure 3. **Regeneration of XLFM light field images via MVM-LF.** The model can still accurately predict the view under appropriate occlusion, indicating that it has learned the global view relationship. Excessive occlusion (90%) causes prediction to crash, indicating that MVM-LF requires a reasonable occlusion ratio to balance information loss and network learning ability.

- Based on XLFM-Former, we propose the Masked View Modeling-Light Field (MVM-LF) self-supervised learning task, which leverages light field viewpoint relationships for masked prediction, improving the model's geometric understanding at a low cost.
- We introduce a PSF-based physical prior loss function, which explicitly models the optical imaging process of light field microscopy, guiding the network to learn reconstruction results that adhere to the optical system's characteristics, thereby enhancing physical consistency and generalization capability.

2. Related Work

2.1. Unsupervised Pretraining Methods for Light Field Microscopy

In the computer vision community, unsupervised pretraining methods have gained widespread attention [3, 13, 14, 21, 25]. For example, Masked Autoencoders [16] learn by randomly masking parts of the input to help the model understand spatial relationships and global context. Contrastive learning [6, 7, 9, 18, 30, 33] creates positive and negative pairs of samples to bring similar samples closer and push dissimilar ones apart. However, in the context of LFM, unsupervised methods are still underexplored. A recent approach [34], Masked LF Modeling (MLFM), introduces a self-supervised pre-training scheme to enhance Light Field Super-Resolution (LFSSR). This method uses a transformer-based structure, LFormer, to learn inter-view correlations. While this approach significantly improves performance, its reliance on random masking does not fully capture the interdependencies between views, which are critical for high-quality super-resolution. Although both approaches are applied to light fields, our proposed method is fundamentally different. First, we focus on the specific task of XLFM reconstruction. Second, our approach is based on viewpoint reconstruction, whereas theirs relies on random pixel masking.

2.2. 3D Reconstruction in XLFM

Recent work [10, 26, 31] has demonstrated the potential of deep learning in addressing computational bottlenecks in XLFM. A recent approach [31] combines two neural networks, SLNet and XLFMNet, for real-time sparse 3D volumetric reconstruction in light field microscopy. SLNet extracts the spatio-temporally sparse components from image sequences, while XLFMNet performs high-fidelity 3D reconstruction. Another recent approach [26] proposes using a conditional normalizing flow architecture for fast 3D reconstruction of neural activity in immobilized zebrafish. However, like XLFMNet, this method remains constrained by its sparsity-driven approach, reconstructing only neural signals while disregarding complete biological morphology. Unlike these prior methods, XLFM-Former is designed for full-volume imaging, reconstructing not only neural activity but also entire volumetric structures. By leveraging a Swin Transformer backbone for hierarchical feature extraction and MVM-LF pretraining to learn global context dependencies, XLFM-Former provides a comprehensive reconstruction of biological samples. This distinction is critical in applications where both functional (neural signals) and anatomical (morphological structures) information are necessary for deeper biological insights.

A recent end-to-end approach [10] combines differentiable simulations of optical systems with deep learning-based reconstruction networks for high-performance computational imaging. The key insight is that global information is crucial for such problems, which is achieved by using Fourier-Nets, a shallow neural network architecture based on global kernel Fourier convolution. However, mapping

to the Fourier domain results in a substantial increase in memory usage, requiring multiple GPUs for large-volume reconstruction. This limits the method's scalability and applicability to more general imaging tasks. This method is particularly expensive in terms of video memory and is not suitable for XLFM reconstruction because the final output of XLFM exceeds 100 million pixels and cannot be made into patches due to system design issues.

To achieve such global information extraction: 1) we use the Swin Transformer [22] as a feature extraction module and then apply a CNN-based decoder for feature fusion. Using self-attention is more efficient than convolution mapped to the Fourier domain. 2)To force the network to understand the dependencies between different views, we propose a proxy task. By masking 70% of the views, we force the network to reconstruct the masked views, enabling unsupervised pretraining.

3. XLFM-Zebrafish Dataset

3.1. Data Collection

To construct the XLFM-Zebrafish Dataset, we utilized an advanced XLFM system designed to capture high-resolution volumetric neural activity in zebrafish. The data collection process was carefully structured to ensure diversity in motion states, imaging conditions, and biological variability. For free-swimming zebrafish, we recorded neural activity in an unconstrained environment, allowing for the study of brain-wide dynamics during naturalistic behaviors. A real-time tracking system was employed to continuously adjust the imaging field of view, ensuring that the zebrafish remained within the microscope's focal range. Additionally, motion artifacts caused by rapid movement were mitigated through dual-color fluorescence imaging and adaptive filtering techniques. In contrast, fixed zebrafish were embedded in a stabilizing medium to facilitate high-precision 3D structural reconstruction. This setting eliminated motion-induced distortions, enabling the extraction of detailed neural architecture. The immobilized specimens were further divided into different groups for training, validation, and testing purposes, ensuring a structured dataset for benchmarking reconstruction algorithms. To capture both rapid neural dynamics and long-term activity trends, we employed multiple imaging conditions that varied in temporal resolution and sampling strategies. This approach allowed us to balance high-fidelity reconstruction with the need for extended observation periods.

3.2. Dataset Statistics

To construct a high-quality XLFM-Zebrafish Dataset, we collected zebrafish in different motion states and set various sampling conditions to ensure diversity and applicability. This dataset is the first standardized XLFM zebrafish 3D reconstruction dataset, designed to evaluate the performance of deep learning models in XLFM-3D reconstruction tasks. The XLFM-Zebrafish Dataset consists of two categories of zebrafish data: Free-swimming Zebrafish, includes 3 individual zebrafish, used for studying dynamic neural activity and analyzing the impact of motion blur on 3D reconstruction. Fixed Zebrafish, includes 13 individual zebrafish, suitable for high-precision 3D structural reconstruction, with 7 individuals for training and validation, and 6 for testing. Additionally, we introduced two different sampling rates:10fps (High sampling rate): Suitable for temporal neural activity modeling and high-precision light field reconstruction. 1fps (Low sampling rate): Used for long-term dynamic tracking and reconstruction stability analysis under low frame rates. The dataset comprises 22,581 light field images. The detailed statistics are presented in the Supplementary Section 7.

4. Methodology

4.1. XLFM-Former

Figure 4 illustrates the overall architecture of XLFM-Former. We describe the details of encoder and decoder in this subsection.

XLFM-Former Encoder: The input to XLFM-Former is a raw light field captured by the XLFM system. The original light field data is represented as a 2D array of sub-aperture views with a resolution of $N \times N$, where N denotes the number of views along both horizontal and vertical directions. Each sub-aperture view contains spatial information about the observed scene, forming a multi-view representation. To align the data with the optical coordinate system and extract relevant information, a cropping operation is applied to each sub-aperture view of cropped light field size $\mathcal{X} \in \mathbb{R}^{H \times W \times D \times S}$. H and W denote the height and width of each cropped sub-aperture view. D represents the depth dimension obtained from the cropping process, which corresponds to the axial resolution. S refers to the number of channels, which is set to $S = 1$ in XLFM light field imaging, producing a 3D view token representation. To facilitate self-attention computation in the transformer-based encoder, the patch partitioning layer divides the input volume into a sequence of non-overlapping 3D View Tokens, each of size (H', W', D'). The resulting tokenized feature map has dimensions: $\frac{H}{H'} \times \frac{W}{W'} \times \frac{D}{D'}$, where each token is projected into a C-dimensional embedding space using a linear embedding layer. This patch-wise tokenization allows the transformer to process spatial and depth information efficiently, enabling effective modeling of local and global dependencies in the light field data. The encoder consists of four hierarchical stages, where each stage contains two consecutive Swin Transformer blocks. At each stage, a patch merging layer is applied to downsample the resolution while

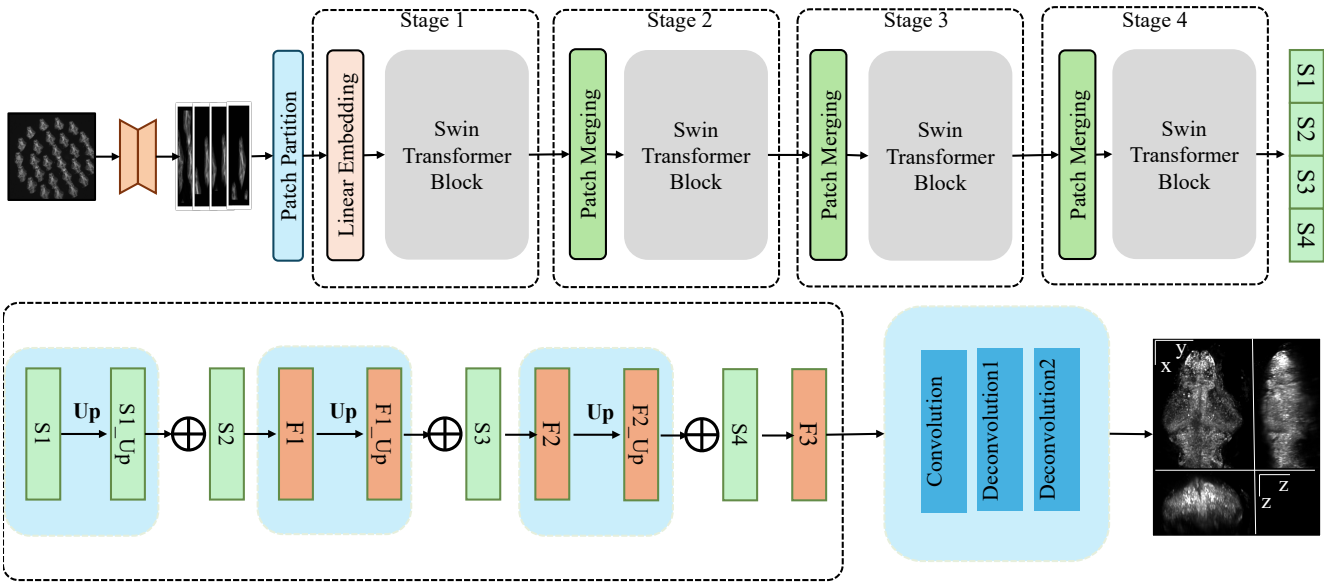

Figure 4. **Overview of the Swin-XLFM architecture.**

increasing feature dimensionality. Given an input feature map $\mathbf{S} \in \mathbb{R}^{h\times w\times d\times c}$ at stage T, the patch merging operation first groups adjacent patches of size $2 \times 2 \times 2$ and concatenates their features where $\mathbf{S}' \in \mathbb{R}^{\frac{h}{2}\times\frac{w}{2}\times\frac{d}{2}\times 8c}$ represents the concatenated feature map. A linear projection layer is then applied to transform the feature dimension to $2c$, resulting in: $\mathbf{S}_{T+1} = \mathbf{W}\mathbf{X}' + \mathbf{b}$, where $\mathbf{W}$ is the learnable weight matrix, and $\mathbf{b}$ is the bias term. Subsequently, the four feature vectors $S_i, \quad i \in \{1, 2, 3\}$ corresponding to the four stages are sent to the decoder for feature fusion.

Decoder: The decoder reconstructs the final high-resolution 3D structure by progressively integrating multi-scale features extracted from the encoder. This process consists of two key components: progressive upsampling and cross-level feature fusion. The decoder takes the hierarchical feature maps from the encoder, denoted as $S_i, \quad i \in \{1, 2, 3\}$, and reconstructs the high-resolution output through a sequence of deconvolutional operations. The highest-level feature map S_1 serves as the starting point, and at each stage, an upsampling operation is applied to gradually recover spatial details. Fused feature maps $F_i, \quad i \in \{1, 2, 3\}$ are obtained by adding features of different scales:

$$\hat{F}_0 = S_1, \quad \hat{F}_i = \mathrm{Up}(\hat{F}_{i-1}) + S_i, \quad i \in \{1, 2, 3\}. \tag{1}$$

The final reconstructed 3D structure is obtained by applying a reconstruction head, which maps the fused feature F_3 to the target 3D volume: First, a convolution is performed to match the number of channels:

$$\hat{F} = \mathrm{Conv}_{1\times1}(F_{\text{fusion}}), \tag{2}$$

Then two deconvolutions are performed to match the target resolution:

$$\mathcal{V}_{\text{pred}} = \mathrm{Deconv}_2\Big(\mathrm{Deconv}_1\big(\mathrm{Conv}_{1\times1}(\hat{F})\big)\Big). \tag{3}$$

4.2. Physics-Guided PSF Reprojection Loss

In order to enforce that the reconstructed volume adheres to the underlying optical physics of the XLFM system, we introduce a physics-guided loss based on the system's Point Spread Function (PSF). The PSF, denoted as h, characterizes the optical blurring inherent in the imaging process.

Given a volumetric reconstruction $\mathcal{V}$ (either the predicted volume $\mathcal{V}_{\text{pred}}$ or the ground truth $\mathcal{V}_{\text{GT}}$), we simulate the corresponding raw light field measurement by convolving the volume with the PSF:

$$\mathbf{I}_{\text{sim}} = h * \mathcal{V}, \tag{4}$$

where $*$ denotes convolution.

Accordingly, we compute the simulated measurements for both the predicted volume and the ground truth:

$$\mathbf{I}_{\text{pred}} = h * \mathcal{V}_{\text{pred}}, \quad \mathbf{I}_{\text{GT}} = h * \mathcal{V}_{\text{GT}}. \tag{5}$$

The Physics-Guided PSF Reprojection Loss is then defined as the mean squared error (MSE) between these two simulated measurements:

$$\mathcal{L}_{\text{PSF}} = \|\mathbf{I}_{\text{pred}} - \mathbf{I}_{\text{GT}}\|_2^2 = \|h * \mathcal{V}_{\text{pred}} - h * \mathcal{V}_{\text{GT}}\|_2^2. \tag{6}$$

By minimizing $\mathcal{L}_{\text{PSF}}$, we encourage the network to generate 3D reconstructions that not only approximate the ground truth in the volumetric domain but also conform to the physical imaging process modeled by the PSF.

4.3. Masked View Modeling for Light Fields (MVM-LF)

To enhance self-supervised learning and inter-view modeling in XLFM, we propose Masked View Modeling for Light Fields (MVM-LF) as a pretraining strategy, enabling the model to reconstruct missing views and capture global scene structures.

Pretrained Encoder: The encoder architecture and input representation in MVM-LF are identical to those used in XLFM reconstruction. This consistency ensures that the learned features during pretraining are directly transferable to the supervised reconstruction task. The pretrained encoder, denoted as f_θ, serves as the initialization for the XLFM reconstruction model.

Lightweight Decoder: Inspired by self-supervised masked reconstruction frameworks, we adopt a lightweight decoder consisting of a series of convolutional layers. The decoder is responsible for predicting the missing views during pretraining. Once training is completed, the decoder is discarded, and only the pretrained encoder is retained for fine-tuning in the XLFM reconstruction task.

Masking Strategy: The core principle of MVM-LF is to randomly mask a proportion r_m of the input views and force the network to reconstruct them based solely on the unmasked views. This proxy task compels the model to learn a joint representation of the global structure and inter-view dependencies. Formally, given a set of sub-aperture views:

$$\mathcal{U} = \{U_1, U_2, \ldots, U_{N_u}\}, \tag{7}$$

we define the masked subset as:

$$\mathcal{U}_{\text{mask}} = \{U_i \mid i \in \mathcal{M}\}, \tag{8}$$

where $\mathcal{M}$ is a randomly sampled index set satisfying $|\mathcal{M}| = r_m N_u$ with $r_m = 0.7$ (i.e., 70% of the views are masked). The network is trained to reconstruct the missing views as:

$$\hat{\mathcal{U}}_{\text{mask}} = f_\theta(\mathcal{U} \setminus \mathcal{U}_{\text{mask}}). \tag{9}$$

The loss function for MVM-LF is defined as the mean squared error (MSE) between the predicted and ground truth masked views:

$$\mathcal{L}_{\text{MVM-LF}} = \sum_{U_i \in \mathcal{U}_{\text{mask}}} \|U_i - \hat{U}_i\|_2^2. \tag{10}$$

4.4. Loss Function

For the pre-training task, we only use ℓ_2 loss. For the XLFM reconstruction task, we complete the reconstruction by minimizing the following loss combination.

$$\begin{aligned}\mathcal{L}_{\text{total}} = \frac{1}{\lambda_1}\mathcal{L}_{\text{MS\_SSIM}} + \frac{1}{\lambda_2}\mathcal{L}_{\text{Edge}} + \frac{1}{\lambda_3}\mathcal{L}_{\text{PSNR}} \\ + \frac{1}{\lambda_4}\mathcal{L}_{\text{MSE}} + \frac{1}{\lambda_5}\mathcal{L}_{\text{PSF}}.\end{aligned} \tag{11}$$

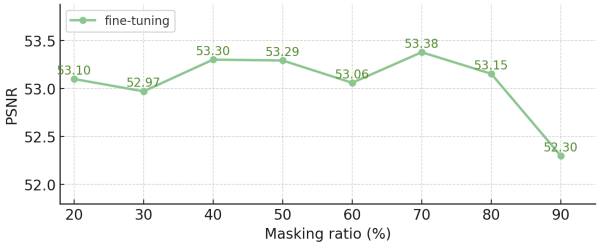

Figure 5. **Masking ratio.**

The detailed loss function presented in the Supplementary Section 8.

5. Experiments

5.1. Implementation Details

For the MVM-LF task, we employ a batch size of 8 to facilitate stable training dynamics. To enhance convergence and mitigate the risk of the model becoming trapped in local optima, we utilize the ReduceLROnPlateau learning rate scheduler, with an initial learning rate set to 1e-4. The training process is conducted for 250 epochs to ensure robust feature learning. All experiments are performed on a distributed computing setup with four NVIDIA A100-80GB SMX4 GPUs. For the XLFM reconstruction task, the training configuration remains largely consistent, with the primary exception that the batch size is set to 1, aligning with the requirements of volumetric reconstruction. The detailed experimental setup presented in the Supplementary Section 9.

5.2. Main Results

We evaluate our approach against state-of-the-art architectures, including ConvNeXt, ViT, PVT, EfficientNet, ResNet-50/101, and U-Net (Table 1). Our method consistently outperforms all baselines across all evaluation metrics, achieving the highest PSNR (54.04 dB) and SSIM (0.9944), significantly surpassing ConvNeXt (50.16 dB, 0.9876) and other transformer- and CNN-based models. Figure 6 further highlights the qualitative advantages of our model, producing sharper, more structurally accurate reconstructions, while competing methods exhibit blurring, distortions, and detail loss. These results demonstrate that physics-guided constraints and hierarchical transformer-based modeling are key to achieving superior XLFM reconstruction, both numerically and perceptually.

5.3. Ablation Study

Masking ratio: As shown in Figure 5, we analyze the impact of different masking ratios in MVM-LF pretraining on XLFM reconstruction. The results indicate that moderate masking (50–70%) achieves optimal performance,

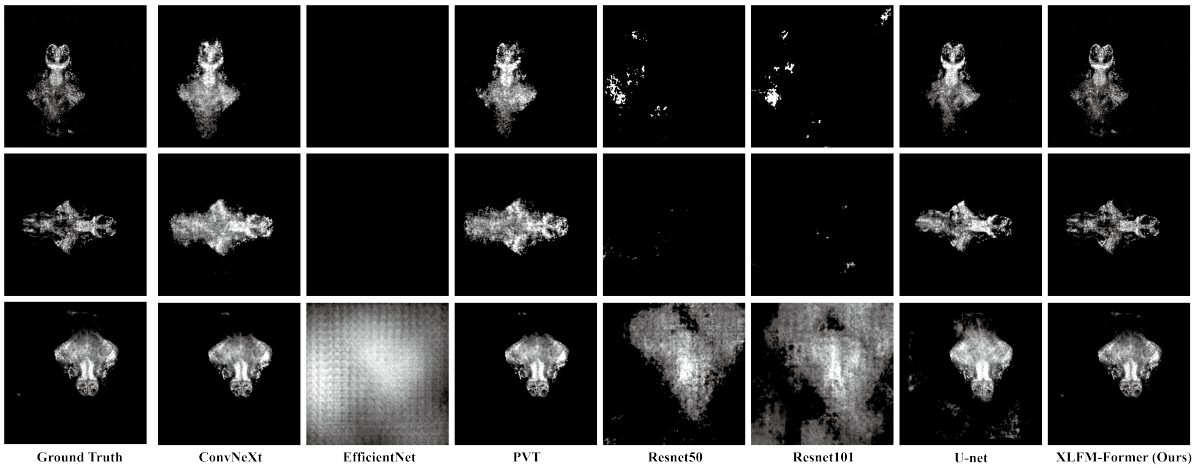

Figure 6. **Compare with the state-of-the-art architecture on XLFM-Zebrafish Dataset.**

Table 1. **Comparison of Methods on XLFM-Zebrafish Dataset.** The best results are highlighted in **bold**, while the second-best are underlined.

	# 1		# 2		# 3		# 4		# 5		# 6		Avg.	
Method	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
ConvNeXt [23]	49.48	0.9851	53.88	0.9867	44.87	0.9833	51.38	0.9882	51.52	0.9892	49.79	0.9935	50.16	0.9876
ViT [12]	49.38	0.9842	52.67	0.9895	45.29	0.9834	51.09	0.9888	51.35	0.9906	45.90	0.9893	49.28	0.9876
PVT [32]	47.21	0.9804	47.93	0.9760	44.50	0.9807	49.46	0.9851	48.32	0.9841	46.60	0.9910	47.34	0.9829
EfficientNet [29]	45.04	0.9550	54.68	0.9851	42.13	0.9541	49.56	0.9801	48.63	0.9772	27.16	0.7264	44.53	0.9296
ResNet-50 [15]	46.46	0.9688	54.89	0.9851	41.46	0.9388	49.47	0.9790	48.82	0.9786	39.98	0.9304	46.85	0.9634
ResNet-101 [15]	47.20	0.9728	54.90	0.9851	41.33	0.9266	49.47	0.9787	49.09	0.9800	39.50	0.8893	46.91	0.9554
U-Net [27]	48.81	0.9807	57.23	0.9928	44.41	0.9808	52.61	0.9908	52.06	0.9904	41.47	0.9725	49.43	0.9847
Ours	**53.97**	**0.9930**	**59.83**	**0.9963**	**49.31**	**0.9910**	**54.55**	**0.9951**	**54.65**	**0.9955**	**51.95**	**0.9956**	**54.04**	**0.9944**

with PSNR peaking at 53.38 dB at 70% masking. Lower masking ratios (20–30%) provide insufficient representation learning, leading to suboptimal fine-tuning results, while excessive masking (90%) reduces PSNR to 52.30 dB, indicating difficulty in reconstructing missing views with limited context. These findings highlight the importance of balancing information removal and reconstruction difficulty, and we adopt 70% masking as the default setting for maximal pretraining efficiency.

Efficacy of Pre-training: We assess data efficiency by comparing models trained from scratch and those with pretraining under varying labeled data proportions (Figure 7). Pretraining provides a significant boost, especially in low-data regimes, with a PSNR of 51.92 dB at 10% labeled data, surpassing the 50.73 dB of training from scratch. While the performance gap narrows as more labeled data is available, the pretrained model consistently outperforms the scratch-trained counterpart, even at 80% labeled data. These results confirm that MVM-LF pretraining enhances feature generalization, making the model more data-efficient and robust across different data availability scenarios.

Efficacy of different components: We conduct an ablation study to assess the impact of PSF loss and MVM-LF pretraining (Table 2). The baseline model achieves 52.14 dB PSNR, while adding PSF loss improves reconstruction fidelity to 52.96 dB. MVM-LF pretraining further enhances global view dependency learning, reaching 53.38 dB. Integrating both components into the full model yields the highest performance (54.04 dB PSNR, 0.9944 SSIM), confirming that combining physics-based constraints with self-supervised pretraining results in the most effective XLFM reconstruction.

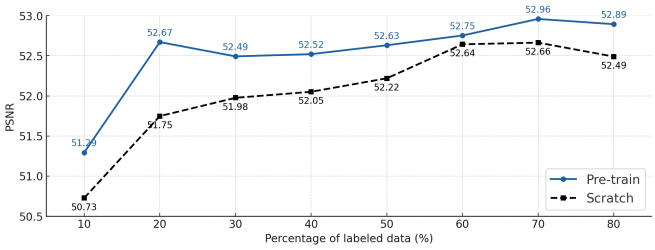

Figure 7. **Data-efficient performance on XLFM-Zebrafish Test Dataset**.

Table 2. **Proposed components.** 'Baseline' refers to the standard XLFM-Former model. 'PSF' indicates the inclusion of the physics-guided PSF loss. 'MVM-LF' corresponds to the incorporation of MVM-LF pretraining. 'Full' represents the complete model with both PSF loss and MVM-LF pretraining. The best results are highlighted in **bold.**

Method	PSNR↑	SSIM↑
baseline	52.1417	0.9924
PSF	52.9560	0.9931
MVM-LF	53.3772	0.9938
Full	**54.0435**	**0.9944**

Table 3. **Pretraining strategies.** 'Baseline' refers to training from scratch. 'ImageNet 1k' and 'ImageNet 22k' indicate models initialized with ImageNet-1k and ImageNet-22k pretrained weights, respectively. 'Random Masked' represents pretraining with a randomly masked pixel reconstruction strategy. 'Ours' corresponds to the proposed MVM-LF pretraining. The best results are highlighted in **bold.**

Method	PSNR↑	SSIM↑
baseline	52.1417	0.9924
ImageNet 1k	52.6976	0.9931
ImageNet 22k	52.3770	0.9923
Random masked	52.9697	0.9934
Ours	**54.0435**	**0.9944**

Table 4. **Missing views in XLFM reconstruction.** The table reports PSNR and SSIM performance when the number of available input views is progressively reduced (from 90% to 60%). The baseline ('Scratch') represents a model trained from scratch with all input views. The results demonstrate that even when only 60% of the views are provided, the pretrained model surpasses the scratch-trained model using full input views, confirming its ability to infer missing views.

Input View Ratio	PSNR↑	SSIM↑
Scratch	52.1417	0.9924
90%	52.9650	0.9933
80%	53.2565	0.9936
70%	52.6687	0.9928
60%	52.5372	0.9928

Efficacy of Pretraining Strategies: We compare MVM-LF pretraining with alternative methods, including ImageNet-based initialization and pixel-level masked pretraining (Table 3). Training from scratch (Baseline) achieves 52.14 dB PSNR, while ImageNet-1k/22k pretraining provides only marginal improvements (52.70 dB / 52.38 dB), indicating that conventional pretraining is suboptimal for XLFM data. Random-masked pretraining performs slightly better (52.97 dB PSNR), but MVM-LF pretraining achieves the best results (54.04 dB PSNR, 0.9944 SSIM), demonstrating its superior ability to model multi-view dependencies. These findings highlight the importance of task-specific pretraining in optimizing XLFM reconstruction quality.

Generalization Under Reduced Input Views: We evaluate the pretrained model's ability to infer missing views by progressively reducing available inputs (Table 4). The scratch-trained model achieves 52.14 dB PSNR with full views, whereas the pretrained model surpasses it even with only 60% of views, reaching 52.54 dB PSNR. The best performance (53.26 dB PSNR) occurs at 80% input views, confirming strong generalization. These results demonstrate that MVM-LF pretraining enables robust multi-view reconstruction, allowing high-fidelity reconstruction even with incomplete input, making it highly adaptable to real-world imaging constraints.

6. Conclusion

We present XLFM-Former, a transformer-based framework for high-speed, high-fidelity XLFM volumetric reconstruction. It integrates a hierarchical Swin Transformer encoder and a physics-guided PSF loss to capture multi-scale spatial features while ensuring optical consistency. We further propose MVM-LF pretraining, a self-supervised strategy that improves global view dependency learning and missing view inference. Experiments on the XLFM-Zebrafish Dataset show that XLFM-Former surpasses SOTA models in PSNR and SSIM. Ablation studies validate the impact of PSF constraints, loss components, and self-supervised pretraining, demonstrating the synergy between physics priors and learned features. Even with reduced input views, MVM-LF pretraining enables robust reconstruction, making XLFM-Former a scalable solution for neuroscience and biomedical imaging.

In the fields of medical imaging and neuroscience, labeled data is often scarce and expensive to obtain, making self-supervised learning an essential direction for future research. Meanwhile, the computer vision community has developed a range of powerful and efficient pretraining strategies and backbone architectures, which have significantly advanced neural imaging techniques. At the same time, novel methodologies emerging from neuroscience have inspired innovations in deep learning model design. We hope that our work serves as a bridge between these two communities, fostering closer interdisciplinary collaboration and encouraging further exploration of self-supervised pretraining for complex biological imaging tasks.

References

[1] Ernst Abbe. Beiträge zur theorie des mikroskops und der mikroskopischen wahrnehmung. *Archiv für mikroskopische Anatomie*, 9(1):413–468, 1873. 2

[2] Lu Bai, Lin Cong, Ziqi Shi, Yuchen Zhao, Yujie Zhang, Bin Lu, Jing Zhang, Zhi-Qi Xiong, Ninglong Xu, Yu Mu, et al. Volumetric voltage imaging of neuronal populations in the mouse brain by confocal light-field microscopy. *Nature Methods*, 21(11):2160–2170, 2024. 1

[3] Amir Bar, Xin Wang, Vadim Kantorov, Colorado J Reed, Roei Herzig, Gal Chechik, Anna Rohrbach, Trevor Darrell, and Amir Globerson. Detreg: Unsupervised pretraining with region priors for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14605–14615, 2022. 3

[4] Edward S Boyden, Feng Zhang, Ernst Bamberg, Georg Nagel, and Karl Deisseroth. Millisecond-timescale, genetically targeted optical control of neural activity. *Nature neuroscience*, 8(9):1263–1268, 2005. 1

[5] Yuming Chai, Kexin Qi, Yubin Wu, Daguang Li, Guodong Tan, Yuqi Guo, Jun Chu, Yu Mu, Chen Shen, and Quan Wen. All-optical interrogation of brain-wide activity in freely swimming larval zebrafish. *iScience*, 27(1):108385, 2024. 1

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 3

[7] Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. *Advances in neural information processing systems*, 33:8765–8775, 2020. 3

[8] Lin Cong, Zeguan Wang, Yuming Chai, Wei Hang, Chunfeng Shang, Wenbin Yang, Lu Bai, Jiulin Du, Kai Wang, and Quan Wen. Rapid whole brain imaging of neural activity in freely behaving larval zebrafish (danio rerio). *elife*, 6:e28158, 2017. 1

[9] Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. Parametric contrastive learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 715–724, 2021. 3

[10] Diptodip Deb, Zhenfei Jiao, Ruth Sims, Alex Chen, Michael Broxton, Misha B Ahrens, Kaspar Podgorski, and Srinivas C Turaga. Fouriernets enable the design of highly non-local optical encoders for computational imaging. *Advances in Neural Information Processing Systems*, 35:25224–25236, 2022. 3

[11] Winfried Denk, James H Strickler, and Watt W Webb. Two-photon laser scanning fluorescence microscopy. *Science*, 248 (4951):73–76, 1990. 1

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*. 7

[13] Dumitru Erhan, Aaron Courville, Yoshua Bengio, and Pascal Vincent. Why does unsupervised pre-training help deep learning? In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 201–208. JMLR Workshop and Conference Proceedings, 2010. 3

[14] Dengpan Fu, Dongdong Chen, Jianmin Bao, Hao Yang, Lu Yuan, Lei Zhang, Houqiang Li, and Dong Chen. Unsupervised pre-training for person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14750–14759, 2021. 3

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7

[16] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 3

[17] Jan Huisken, Jim Swoger, Filippo Del Bene, Joachim Wittbrodt, and Ernst HK Stelzer. Optical sectioning deep inside live embryos by selective plane illumination microscopy. *Science*, 305(5686):1007–1009, 2004. 1

[18] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020. 3

[19] Marc Levoy, Ren Ng, Andrew Adams, Matthew Footer, and Mark Horowitz. Light field microscopy. In *Acm siggraph 2006 papers*, pages 924–934. 2006. 1

[20] Haoyu Li, Changliang Guo, Deborah Kim-Holzapfel, Weiyi Li, Yelena Altshuller, Bryce Schroeder, Wenhao Liu, Yizhi Meng, Jarrod B French, Ken-Ichi Takamaru, et al. Fast, volumetric live-cell imaging using high-resolution light-field microscopy. *Biomedical optics express*, 10(1):29–49, 2018. 1

[21] Hao Liu and Pieter Abbeel. Behavior from the void: Unsupervised active pre-training. *Advances in Neural Information Processing Systems*, 34:18459–18473, 2021. 3

[22] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 4

[23] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 7

[24] Leon B Lucy. An iterative technique for the rectification of observed distributions. *Astronomical Journal, Vol. 79, p. 745 (1974)*, 79:745, 1974. 2

[25] Oscar Manas, Alexandre Lacoste, Xavier Giró-i Nieto, David Vazquez, and Pau Rodriguez. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9414–9423, 2021. 3

[26] Josué Page Vizcaíno, Panagiotis Symvoulidis, Zeguan Wang, Jonas Jelten, Paolo Favaro, Edward S Boyden, and Tobias Lasser. Fast light-field 3d microscopy with out-of-distribution detection and adaptation through conditional normalizing flows. *Biomedical optics express*, 15(2):1219–1232, 2024. 3

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 7

[28] C. F. Shang, Y. F. Wang, M. T. Zhao, Q. X. Fan, S. Zhao, Y. Qian, S. J. Xu, Y. Mu, J. Hao, and J. L. Du. Real-time analysis of large-scale neuronal imaging enables closed-loop investigation of neural dynamics. *Nat Neurosci*, 27(5):1014–1018, 2024. 1

[29] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 7

[30] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in neural information processing systems*, 33:6827–6839, 2020. 3

[31] Josue Page Vizcaino, Zeguan Wang, Panagiotis Symvoulidis, Paolo Favaro, Burcu Guner-Ataman, Edward S Boyden, and Tobias Lasser. Real-time light field 3d microscopy via sparsity-driven learned deconvolution. In *2021 IEEE International Conference on Computational Photography (ICCP)*, pages 1–11. IEEE, 2021. 3

[32] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 568–578, 2021. 7

[33] Xiao Wang and Guo-Jun Qi. Contrastive learning with stronger augmentations. *IEEE transactions on pattern analysis and machine intelligence*, 45(5):5549–5560, 2022. 3

[34] Da Yang, Hao Sheng, Sizhe Wang, Shuai Wang, Zhang Xiong, and Wei Ke. Boosting light field spatial super-resolution via masked light field modeling. *IEEE Transactions on Computational Imaging*, 2024. 3

[35] Feng Zhang, Li-Ping Wang, Edward S Boyden, and Karl Deisseroth. Channelrhodopsin-2 and optical control of excitable cells. *Nature methods*, 3(10):785–792, 2006. 1

[36] Zhenkun Zhang, Lu Bai, Lin Cong, Peng Yu, Tianlei Zhang, Wanzhuo Shi, Funing Li, Jiulin Du, and Kai Wang. Imaging volumetric dynamics at high speed in mouse and zebrafish brain with confocal light field microscopy. *Nature biotechnology*, 39(1):74–83, 2021. 1

XLFM-Former: Rapid 3D Reconstruction for Light Field Microscopy

Supplementary Material

7. Dataset Statistics Details

Table 5 summarizes the statistics of the XLFM-Zebrafish dataset, which consists of image data from zebrafish in both free-swimming and fixed conditions. The dataset is categorized into Pre-training Set, Training/Validation Set, and Test Set to facilitate different stages of model development. The dataset is carefully designed to ensure diversity in motion complexity, viewpoint variations, and temporal resolutions. By pretraining on large-scale, complex free-swimming zebrafish data, the model gains a stronger ability to generalize and better reconstruct simpler fixed zebrafish data, leading to improved performance. This structured dataset and training methodology provide a standardized benchmark for evaluating XLFM-Former and contribute to the advancement of XLFM-based 3D reconstruction techniques.

Pre-training Set (Free-swimming Zebrafish): The Pre-training Set comprises three free-swimming zebrafish (m1, m2, m3), totaling 20,123 images, all captured at a 10 fps sampling rate. Free-swimming zebrafish exhibit highly dynamic and complex motion, leading to significant variations in viewpoint and pose. This complexity presents a challenge for 3D reconstruction but also offers an opportunity for the model to learn richer geometric structures. To leverage this complexity, we employ unsupervised pretraining on this large-scale free-swimming dataset before training on the fixed zebrafish dataset. By learning from diverse, naturally occurring light field transformations, the model develops a robust understanding of light field geometry and depth relationships. This approach significantly improves performance when fine-tuned on simpler fixed zebrafish data, demonstrating the benefits of pretraining on large, diverse datasets before supervised learning on smaller, more controlled datasets.

Training/Validation Set (Fixed Zebrafish): The Training/Validation Set contains seven fixed zebrafish (f1-f7) with a total of 1,761 images. Most samples were collected at 10 fps, while some (e.g., f3) were acquired at 1 fps to introduce variations in temporal resolution. Compared to free-swimming zebrafish, the fixed zebrafish dataset presents a more structured and constrained setting, making it an ideal target for supervised training once the model has been pretrained on more complex free-swimming data.

Test Set (Fixed Zebrafish): The Test Set consists of six fixed zebrafish (t1-t6) with a total of 1,011 images. Some samples (e.g., t2, t4, t5) were acquired at 1 fps, allowing a comprehensive evaluation of XLFM-Former's reconstruction performance under different sampling conditions.

Table 5. **The XLFM-Zebrafish dataset statistics.**

Dataset Name	Number of Images	Sampling Rate (fps)
Free-swimming Zebrafish (Pre-training Set)		
moving_fish1 (m1)	4000	10
moving_fish2 (m2)	7332	10
moving_fish3 (m3)	8791	10
Fixed Zebrafish (Training/Validation Set)		
fixed_fish1 (f1)	240	10
fixed_fish2 (f2)	117	10
fixed_fish3 (f3)	318	1
fixed_fish4 (f4)	314	10
fixed_fish5 (f5)	374	10
fixed_fish6 (f6)	214	10
fixed_fish7 (f7)	184	10
Fixed Zebrafish (Test Set)		
test_fixed_fish1 (t1)	300	10
test_fixed_fish2 (t2)	41	1
test_fixed_fish3 (t3)	334	10
test_fixed_fish4 (t4)	61	1
test_fixed_fish5 (t5)	61	1
test_fixed_fish6 (t6)	214	10

8. Loss Function Details

In our proposed XLFM-Former framework, we employ different loss functions tailored for **pretraining** and **reconstruction tasks** to ensure robust and high-quality 3D volume generation.

Pretraining Loss: For the self-supervised pretraining task, where we use Masked View Modeling-Light Field (MVM-LF), we adopt a simple ℓ_2 loss to enforce the consistency between the predicted and ground truth light field views:

$$\mathcal{L}_{\text{pretrain}} = \|\hat{I} - I\|_2^2, \tag{12}$$

where $\hat{I}$ represents the predicted light field views, and I denotes the original (ground truth) views before masking. The choice of ℓ_2 loss is motivated by its stability in regression tasks and its ability to ensure smooth reconstructions during pretraining.

XLFM Reconstruction Loss: For the final **3D reconstruction task**, we minimize the following composite loss function:

$$\begin{aligned}\mathcal{L}_{\text{total}} = &\frac{1}{\lambda_1}\mathcal{L}_{\text{MS\_SSIM}} + \frac{1}{\lambda_2}\mathcal{L}_{\text{Edge}} + \frac{1}{\lambda_3}\mathcal{L}_{\text{PSNR}} \\ &+ \frac{1}{\lambda_4}\mathcal{L}_{\text{MSE}} + \frac{1}{\lambda_5}\mathcal{L}_{\text{PSF}}.\end{aligned} \tag{13}$$

Each term in the loss function contributes to a different aspect of the 3D reconstruction quality, ensuring sharpness,

accuracy, and optical consistency. Below, we provide a detailed breakdown of each component(PSF loss has been described in Section 4.2.):

Multi-Scale Structural Similarity (MS-SSIM) Loss: Structural Similarity Index (SSIM) is widely used to measure the perceptual similarity between images. We employ a multi-scale SSIM (MS-SSIM) loss to capture both local and global structural fidelity:

$$\mathcal{L}_{\text{MS\_SSIM}} = 1 - \text{MS-SSIM}(\hat{V}, V), \tag{14}$$

where $\hat{V}$ and V represent the predicted and ground truth 3D volumes. MS-SSIM helps preserve structural details and enhances the perceptual quality of the reconstruction.

Edge-Aware Loss: To enhance edge sharpness and suppress blurring, we define the edge-preserving loss as a weighted combination of edge loss and multi-scale MSE loss:

$$\begin{aligned}\mathcal{L}_{\text{Edge\_Aware}} = \frac{1}{\lambda_{\text{edge}}}\Big(&\|\nabla_x\hat{V} - \nabla_x V\|_1 + \|\nabla_y\hat{V} - \nabla_y V\|_1 \\ &+ \|\nabla_z\hat{V} - \nabla_z V\|_1\Big) \\ &+ \frac{1}{\lambda_{\text{mse}}}\sum_{s=1}^{S}\|\hat{V}^{(s)} - V^{(s)}\|_2^2.\end{aligned} \tag{15}$$

Here, $\mathcal{L}_{\text{edges\_loss}}$ ensures sharpness by penalizing gradient differences along spatial dimensions, while $\mathcal{L}_{\text{multi\_scale\_MSE}}$ enforces consistency across multiple resolutions, improving global structure and fine details.

Peak Signal-to-Noise Ratio (PSNR) Loss: PSNR is a widely used metric for measuring signal fidelity. We define a loss function that penalizes low PSNR values:

$$\mathcal{L}_{\text{PSNR}} = -\text{PSNR}(\hat{V}, V), \tag{16}$$

where a higher PSNR corresponds to higher-quality reconstructions. By minimizing this loss, we encourage the model to reduce noise and artifacts.

Mean Squared Error (MSE) Loss: The **MSE loss** ensures pixel-wise intensity similarity between the predicted and ground truth volumes:

$$\mathcal{L}_{\text{MSE}} = \|\hat{V} - V\|_2^2. \tag{17}$$

While MSE is commonly used in image restoration, it is prone to blurring. Thus, we use it in combination with perceptual losses (e.g., MS-SSIM and edge loss) to balance fine details and overall similarity.

8.1. Ablation Study on Loss Functions

To evaluate the impact of different loss components on the overall performance of our model, we conduct an ablation study by removing individual loss terms and measuring the performance in terms of PSNR and SSIM. The results are summarized in Table 6.

Table 6. **Loss function.** 'Full' denotes the complete model with all loss terms. 'w.o. ms_ssim' removes the multi-scale SSIM loss. 'w.o. Edge_Aware' omits the edge-aware loss. 'w.o. PSNR' excludes the PSNR loss. 'w.o. MSE Loss' removes the MSE loss. The best results are highlighted in **bold.**

Method	PSNR↑	SSIM↑
w.o. $\mathcal{L}_{MS\_SSIM}$	53.2787	0.9937
w.o. $\mathcal{L}_{Edge\_Aware}$	52.1870	0.9922
w.o. $\mathcal{L}_{PSNR}$	53.0741	0.9935
w.o. $\mathcal{L}_{MSE}$	53.2521	0.9937
Full	**54.0435**	**0.9944**

From the table, we observe that the complete model (*Full*) achieves the highest PSNR of **54.0435** and SSIM of **0.9944**, demonstrating the effectiveness of integrating all loss terms.

Effect of Multi-Scale SSIM Loss: Removing the multi-scale SSIM loss (*w.o.* $\mathcal{L}_{MS\_SSIM}$) results in a performance drop, reducing PSNR to 53.2787 and SSIM to 0.9937. This highlights the importance of SSIM-based perceptual loss in preserving structural information.

Effect of Edge-Aware Loss: When the edge-aware loss is removed (*w.o.* $\mathcal{L}_{Edge\_Aware}$), the PSNR decreases significantly to 52.1870, while SSIM drops to 0.9922. This indicates that the edge-aware term plays a crucial role in maintaining sharp details and edge consistency.

Effect of PSNR Loss: Excluding the PSNR-based loss (*w.o.* $\mathcal{L}_{PSNR}$) results in a minor degradation in performance, with PSNR reducing to 53.0741 and SSIM to 0.9935. This suggests that optimizing directly for PSNR provides some benefits but is not the dominant factor.

Effect of MSE Loss: When the MSE loss is removed (*w.o.* $\mathcal{L}_{MSE}$), the PSNR slightly drops to 53.2521, while SSIM remains at 0.9937. This indicates that MSE contributes to pixel-wise accuracy but is less critical than the other loss terms.

9. Training Configuration

For the MVM-LF task, we adopt a batch size of 8 to ensure stable training dynamics while maintaining an efficient balance between computational cost and convergence stability. To further enhance the training process and prevent the model from becoming trapped in local optima, we employ the ReduceLROnPlateau learning rate scheduler. The initial learning rate is set to 1e-4, and the scheduler dynamically adjusts the learning rate based on validation loss fluctuations, reducing it when no improvement is observed for a predefined number of epochs. This adaptive learning rate strategy helps in maintaining a steady convergence while

Table 7. Data augmentation techniques applied during training. Each transformation is applied with a probability of 0.5.

Augmentation Type	Probability
Random Flip	0.5
Random Rotation	0.5
Random Gaussian Noise&Blur	0.5
Random Brightness &Contrast	0.5

preventing premature stagnation in suboptimal solutions.

To achieve robust feature learning and ensure generalization across diverse data distributions, we train the model for 250 epochs. Given the complexity of the task and the high-dimensional nature of the input data, prolonged training allows the model to capture intricate spatial and structural information effectively. All experiments are conducted in a distributed computing environment equipped with four NVIDIA A100-80GB SMX4 GPUs, leveraging multi-GPU parallelism to accelerate training and optimize resource utilization.

For the XLFM reconstruction task, the training setup remains largely consistent with the MVM-LF configuration. However, a notable difference is the use of a batch size of 1, which aligns with the requirements of volumetric reconstruction. Given that volumetric data often involves higher memory footprints due to its three-dimensional representation, a smaller batch size ensures that computations remain feasible within available GPU memory constraints. All experiments are implemented using PyTorch Lightning, a high-level deep learning framework that simplifies training and enhances reproducibility.

9.1. Data Augmentation Strategy

To improve the model's robustness and generalization capability, we apply a series of data augmentation techniques during training. These augmentations are applied with a probability of 0.5, as summarized in Table 7.

These augmentation techniques enhance the model's ability to handle variations in real-world data, reducing overfitting and improving generalization performance.

9.2. Model Architecture

For all experiments, we adopt the Swin Transformer framework in its tiny configuration. The Swin Transformer is a hierarchical vision transformer that efficiently models long-range dependencies while maintaining computational efficiency. The tiny variant provides a lightweight architecture suitable for our training setup, balancing performance and efficiency.

9.3. Extended Training on Additional Datasets

To further explore the scalability and generalization capability of XLFM-Former, we conducted large-scale training on an extended dataset. This additional training aimed to improve the model's ability to reconstruct fine-grained details while enhancing robustness across diverse volumetric data.

The large-scale training process required substantial computational resources, consuming approximately 1344 A100-80GB GPU hours. This extensive training allowed the model to refine its feature representation and leverage a broader data distribution, leading to improved reconstruction quality.

To demonstrate the effectiveness of this large-scale training, we provide **a visualized demo** showcasing the enhanced view synthesis capability. The qualitative results highlight the model's ability to generate high-fidelity reconstructions with improved structural consistency and perceptual quality, further validating the benefits of extended training.