High-fidelity, high-resolution light-field reconstruction with physics-informed neural representation learning

Chengqiang Yi[1], Yifan Ma[1], Xinyue Yuan[1], Lanxin Zhu[1], Jiahao Sun[1], Shangbang Gao[3], Meng Zhang[4], Yuhui Zhang[4], Zhaoqiang Wang[5], Hsiai Tzung[5], Dongyu Li[1,2], Binbing Liu[1*], Peng Fei[1,2*]

[1]*School of Optical and Electronic Information-Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, Wuhan, China, 430074*

[2]*Advanced Biomedical Imaging Facility Huazhong University of Science and Technology, Wuhan, Hubei 430074, China*

[3] *College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, 430074, China.*

[4]*Britton Chance Center and MoE Key Laboratory for Biomedical Photonics, Wuhan National Laboratory for Optoelectronics-Huazhong University of Science and Technology, Wuhan, Hubei 430074, China*

[5]*Department of Bioengineering, Henry Samueli School of Engineering and Applied Science, University of California, Los Angeles, Los Angeles, 90095, U.S.A.*

[6]*Division of Cardiology, Department of Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, 90095, U.S.A*

[*] Correspondence：liubinbing@hust.edu.cn and feipeng@hust.edu.cn

Abstract:

Fourier light field microscopy (FLFM) has made a name for itself by providing high-speed volumetric imaging of dynamic processes through 2D light field snapshot. However, the inverse reconstruction in FLFM, characterized by a limited-angle tomography 3D reconstruction challenge, suffers the artifacts and axial elongation. Recently, the implicit neural representation (INR) has emerged as a powerful paradigm in 3D reconstruction due to its continuous signal representation, facilitating arbitrary views synthesis with geometry consistency. In this study, we report a novel INR-based FLFM 3D reconstruction method, termed LF-INR, that leverage the inherent continuity from the inductive bias of INR to eliminate the defects of sparse-view reconstruction and enhance the structural fidelity in a self-supervised learning manner. Additionally, an optimized sampling strategy and transfer-learning are adopted to mitigate the internal limitations of conventional INR approach in generalization and optimization speed. We demonstrate its utility on diverse biological simulation data, which shows LF-INR achieves artifacts-free 3D reconstruction and ~1.4 folds structural fidelity improvement compared with deconvolution approach, as well as ~10 folds computational acceleration than conventional INR-based model.

Introduction

Elucidating how the spatial dynamics and interactions in microenvironments are mapped onto the functional representation requires high spatiotemporal resolution across three-dimension. It promotes the development of various volumetric imaging technology to interrogate the fast biological processes, like cardiac hemodynamic[1, 2], intracellular interaction[3-5] etc. Of these, scanning-based microscopy makes compromise on volumetric imaging speed and would even induce extra phototoxicity because of its sequential scanning acquisition strategy. To fulfill the utilization of emission photons

in one excitation, light field microscopy [6] encodes all fluorescence signals from whole excited volume to one 2D image in a term of multi-view captures. This one-shot volumetric imaging capability facilitates the wide application of LFM on dynamic bioprocess[7]. However, the non-uniform and low spatial resolution brought by the frequency aliasing also influence its further application. To mitigate this degradation, by employing the microlens array at Fourier plane, Fourier light field microscopy [8, 9] overcomes the spatial sampling aliasing in standard LFM and can better leverage the optical aperture to achieve improved spatial resolution with less views. These approaches mostly adopted Richard Lucy deconvolution to iteratively reconstruct the 3D volume with point spread function. However, under the lower angular sampling rate and limited angle-range of FLFM, this alternate projection algorithm in a tomography reconstruction manner will introduce artifacts and influence structure fidelity especially for dense signal[10, 11].

To bypass the internal limitations in multi-view 3D reconstruction, data-driven based algorithms are proposed to solve this inverse problem. Building a deep learning model for converting the 2D LF captures to high-resolution 3D stacks obviously improve the reconstruction quality[12-15]. However, these end-to-end supervised learning approaches require large-scale high-quality volumetric data, and the well-aligned "2D-3D" datasets are hard to be obtained for capturing fast 3D dynamics. Besides, the mismatch between training data and experimental data would introduce undesirable artifacts. So, a self-supervised learning-based algorithm without need of volumetric data would be a choice for 3D reconstruction.

Current self-supervised approaches in microscopic vision tasks mostly abide the base principle that incorporating the physical prior into models. Specifically, well-designed loss function derived from analytical model is used to make the network prediction physical-consistent[16, 17]. As for 3D reconstruction inverse problem, an apparent way to construct a self-supervised model based on such architectures by forward projection model constrains. However, the dimensionality reduction in 3D reconstruction inverse problem exacerbates the complexity to search a reasonable solution among the latent solution space under the physical consistency, manifested as the severe hallucinations and discontinuity (Supplementary Figure 1). Therefore, a big concern in self-supervised 3D is how to ensure the continuity of 3D reconstructed signals under physical consistency constrains alone.

Recently, Neural Radiance Fields[18] (NeRF), featured in generating novel views with geometry consistency, has made wide application on various vision tasks, like photo-editing[19], 3D surface extraction and reconstruction[20], human avatar modelling[21] and novel view synthesis[22]. It adopts an MLP (Multi-Layer Perceptron) to represent 3D scene from 2D multi-view input, called INR (implicit neural representation), which is equivalent to obtain the latent continuous function of signal distribution. This volumetric representation is highly correlated with 3D reconstruction inverse problem in FLFM. However, the forward model between macron photography and diffraction theory in microscopy imaging is completely different. Besides, the grid-based training strategy in NeRF burdens the computational cost and influence its further application

Inspired by this, we proposed a physical-informed self-supervised light field 3D reconstruction approach, termed LF-INR, which firstly combines the implicit neural representation with light-field forward model and achieves 3D reconstruction under 2D views' supervised learning. Due to the continuity of INR function and geometry consistency of synthesized views, LF-INR outperforms in reducing artifacts and axial elongation under limited view input, compared with traditional analytical-model based approach. Secondly, we presented "coarse to fine" and 3D spatial decomposition training strategy to reduce the computational time cost compared with conventional 3D grid-based INR. Besides, considering the intrinsic continuity of dynamic LF time-sequence, we adopted network transfer-learning method to accelerate the 3D reconstruction for time-sequential captures. We validated the artifacts removal and axial elongation suppression of LF-INR on dense signals, and demonstrate its limited-view reconstruction ability compared with traditional approach. With its high-fidelity reconstruction, LF-INR improves the accuracy of common biological downstream analysis.

Results

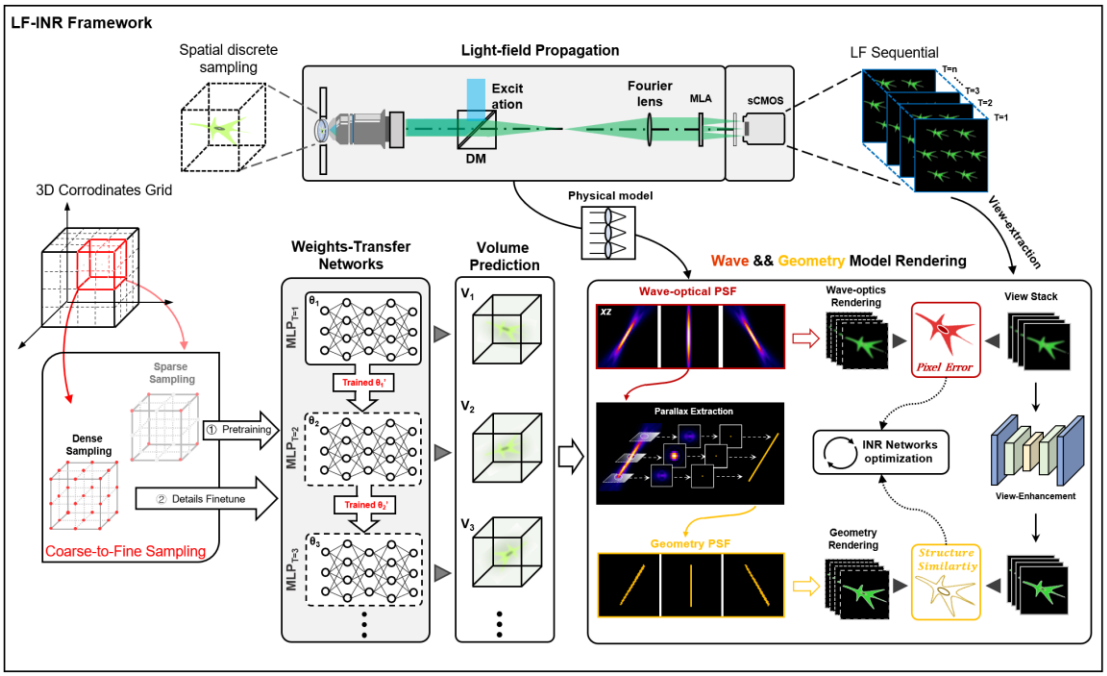


Fig1. The overview of LF-INR. LF-INR adopts multi-view LF images as groudtruth to guide the neural networks to represent the 3D volume. The pre-defined corrdinates grid ("3D Corrordinates Grid ") were sent to MLPs to generate predicted 3D images. With physical-informed rending model ("Wave && Geometry Model Rendering"), muli-view inference can be obtained and then were used to calculate the specific loss function with input LF captures. After one network (such as $MLP_{T=1}$) achieved well-convergence, the optimized parameters were assigned to the next network ($MLP_{T=2}$ as shown in figure) to accelerate 3D reconstruction from sequential LFs.

The principle of LF-INR

The framework of LF-INR includes 4 parts: 3D coordinates grid construction, network inference, forward model rendering and accelerated sequential 3D reconstruction. Unlike conventional CNN-based networks in image enhancement or reconstruction, INR requires spatial coordinates $\{x_i, y_i, z_i\}_{i=1}^{N}$ as the network input, which means a meshgrid $(x - y - z)$ covering the reconstructed signals needed to be defined before network inference. However, this coordinate-wise-based optimization leads to slow network convergence speed, especially multiplicatively increased the time consumption with the enlargement of reconstruction volume. Here, inspired by the two-stage sampling strategy[18] , we introduced a coarse-to-fine sampling strategy, which first used down-sampled meshgrid to pre-train network and next fed all grids for details fitting, to make network fast convergence (as Fig. 1 "Coarse to Fine Sampling" shows). Besides, to reduce the computational operations decided by the coordinate's quantity, LF-INR decomposed the 3D coordinates representation $(x - y - z)$ to two separate components $(xy\text{-}z)$ in consideration of the uneven sampling rate and resolution between lateral and axial planes (Methods, Supplementary Fig. 2). With this two-step sampling strategy and decomposed-representation, LF-INR achieves ~3 folds acceleration than conventional INR optimization approach (Supplementary Fig. 3).

After the coordinates sampling, an MLP was built to learn the mapping function from sampled coordinates to their corresponding values. For example, as shown in "Weight-Transfer Networks", $\text{MLP}_{T=1}$ was used to mapping the selected coordinates to the predicted volume V_1. Then, this volume was used to generate projected views by forward rendering model. However, there existed difference between macron photography rendering model in conventional Nerf and the forward model in microscopy imaging. The former is based on ray tracing to form the volumetric density of opaque 3D scene. While the latter is based on wave-optics theory and aims to reconstruct the intensity distribution of semi-/transparent sample. The extra degradation induced by diffraction will influence the accuracy of geometry-based rendering model. Take this into account, we combined the rendering model of *vanilla* NeRF and light-field PSF (point spread function) to achieve dual forward projection during training phase (as Fig1 "Wave && Geometry Model Rendering" shows). The LFPSF containing optical blur degradation was used to convolved with predicted volume to mimic the light-field propagation decided by hardware settings. In this rendering branch, we adopt pixel-wise loss (L_2 loss) to update the INR parameters. For the second rendering branch, via parallax extraction, the geometry relation between volume and captured views can be first obtained (Supplementary Note 1). Then, we use pre-trained view-enhancement self-supervised network to enhance the resolution of input views (Supplementary Note 2), which is coincided with the geometry rendering model. Accordingly, we use structural loss to capture the skeleton information. Compared with one rendering strategy, our hybrid projection model achieves fast network convergence and high-structure fidelity (Supplementary Fig. 4).

For sequential LF captures, the first frame was used to act self-supervised 3D reconstruction with above accelerated optimization process and the reconstruction time of remained frames were further reduced by network transfer. As shown in ""Weight-Transfer Networks", the optimized parameters of $\text{MLP}_{T=1}$ were used to initialized the

network $MLP_{T=2}$ during the reconstruction process of second LF frame (Methods). This initialized parameters promotes the netowork covergence due the instrisic conituty of biological dynamics and achieved about ~11 folds acceleration compared with the optimization process of conventional NeRF (Supplementary Fig. 3). The whole training process of sequential LF captures is detailed in Supplementary Note 3.

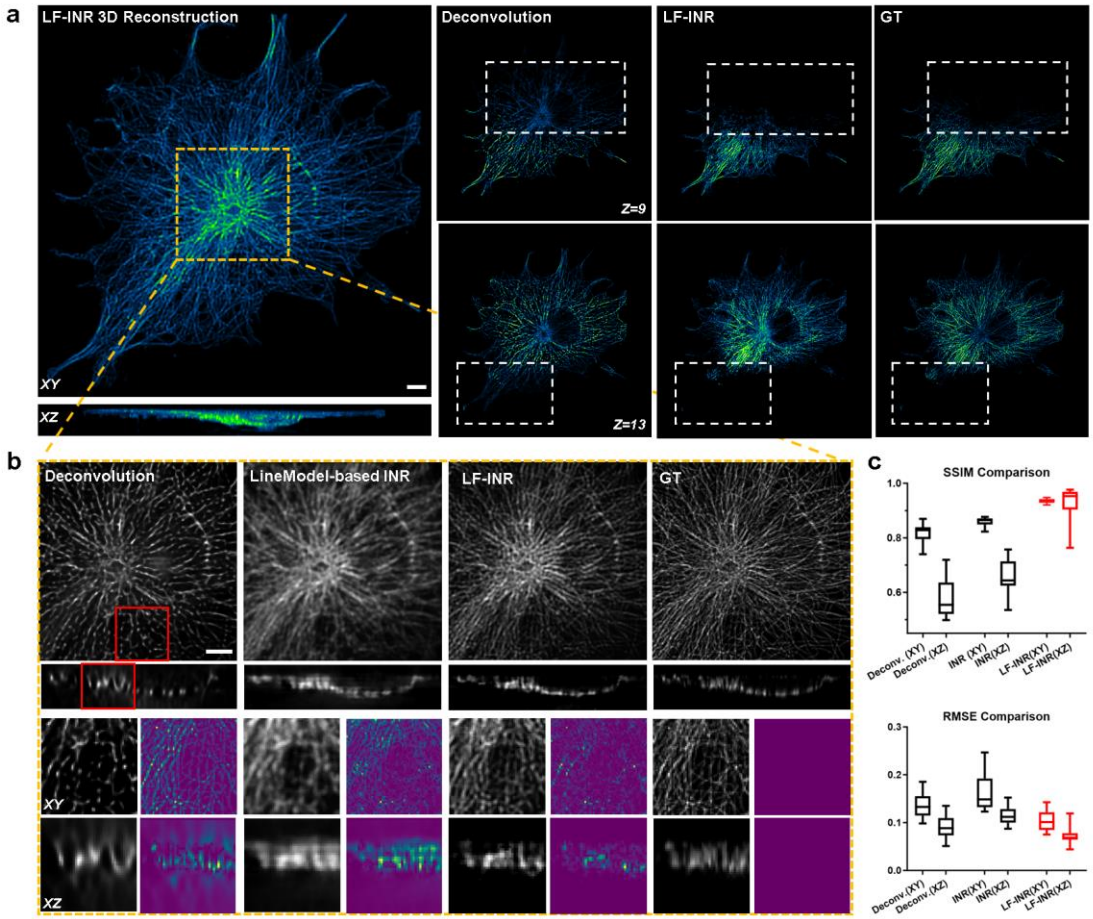


Fig.2 Comparative performances of deconvolution, Geometry-Model-Based INR and LF-INR on the reconstruction of microtube. a, The comparison between deconvolution results and LF-INR at different depth. The white rectangle denoted the artifacts removal by LF-INR. Scale bar: 5 μm. **b,** Higher magnification view of yellow rectangular region marked in (a), denoting the higher structure fidelity of LF-INR compared with deconvolution and Geometry-Model-Based INR. Scale bar: 5 μm. **c,** The statistical results of SSIM and RMSE of different approaches.

Axial improvement of 3D reconstruction offered by LF-INR.

For the inverse problem of 3D reconstruction from multi-view captures, the most widely used approach is RL-deconvolution. Though this approach achieves 3D reconstruction with improved resolution compared with input views in the absence of volumetric data guidance, the intrinsic limitation of this inverse problem (limited angle range and sampling rate of captured views) still influences the reconstruction quality along axial especially for dense signals. To demonstrate the unique features of LF-INR, we conduct reconstruction comparison between deconvolution method and INR-based approaches as shown in Fig 2. We simulated one FLMF image (7 views) from 3D

microtube image (Methods). As Fig. 2a depicted, LF-INR shows less signal crosstalk from other depths compared with deconvolution. In spite of the visual outperformance, the reconstruction quality of dense signal further illustrates the higher sectioning ability brought by implicit neuron representation (Fig. 2b). The reconstructed signals of INR-based approaches (Geometry-Model-Based INR && LF-INR) act better concentrated along Z axis than deconvolution one and retain the continuity and structure similarity as shown in lateral projection. We attributed this improved axial resolution to the novel view synthesis ability and spatial continuity of INR (Supplementary Figure 5). It's noteworthy that because the Geometry-Model-Based INR only used the ray tracing rendering model without the consideration of blurry degradation brought by diffraction effect, the RMSE (root mean square error) is higher than deconvolution approach. While LF-INR inherited the better sectioning ability of INR and adopted the hybrid rendering model to reduce both the pixel error and structural distortion as the statistical results demonstrated in Fig. 2c.

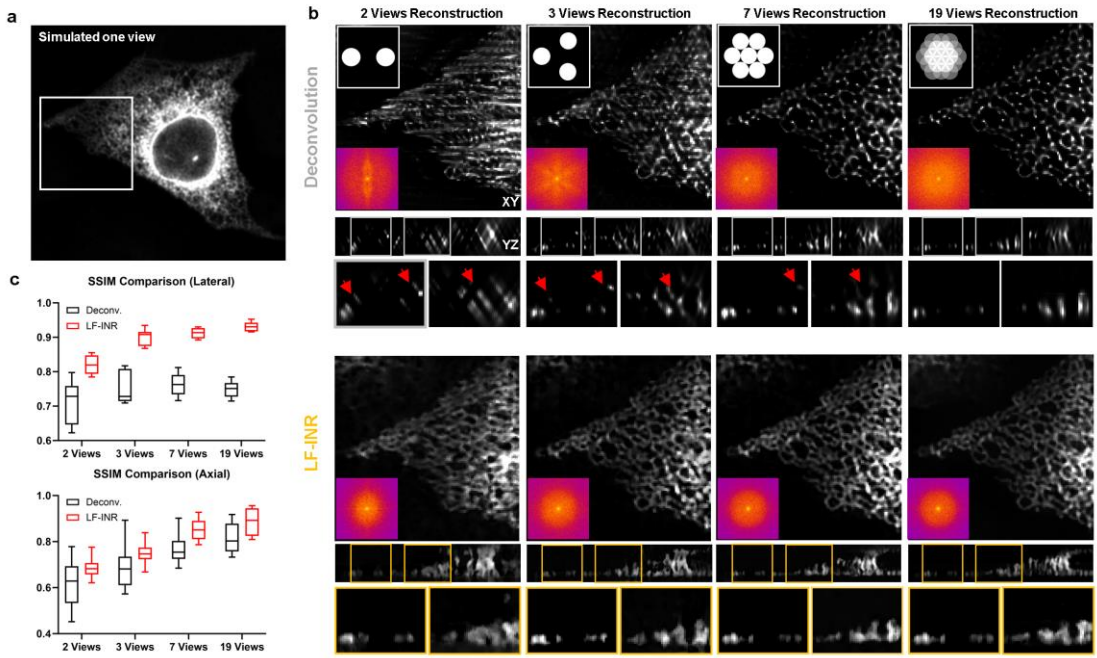


Fig.3 LF-INR enables artifacts suppression on 3D reconstruction from few-view captures. a, One simulated view of ER (endoplasmic reticulum) via FLFM forward projection with pre-defined angle. **b,** Reconstruction comparison via deconvolution and LF-INR under different numbers of input views. Row indicate the lateral projection/axial slice of deconvolution results and LF-INR prediction. Columns indicate the 3D reconstruction under different views. **c,** The statistical results of SSIM.

Artifacts removal in limited-views 3D reconstruction offered by LF-INR.

The principle underlying 3D reconstruction from multi-views capture involves the task of decoding the signal distribution derived from a finite number of views captured within a limited angular range. The higher angular sampling rate benefits the resolvability on dense signal and suppress axial artifacts while the angular range

influences the axial elongation of decoded signals[10]. As for FLFM imaging settings, the existed tradeoff between angular sampling rate and spatial resolution impedes the multi-view reconstruction quality. Increase the limited “angular-spatial” bandwidth without extra sacrifice would promote the reconstruction quality similar. Inspired by the powerful ability of novel views synthesis of NeRF, we conducted the experiments to investigate the robustness of deconvolution approach and LF-INRs under different numbers of input views (Fig 2). To make the fair comparison in the same area of reconstruction results, we generated arbitrary number of FLFM views within the same resolution. The microlens slid in the fixed optical pupil, which yielded 2 & 3 & 7 & 19 FLFM views (see Methods). As Fig 2b shows, deconvolution results exhibited severe striped artifacts along axial plane (as red arrowheads depicted), when only 2 & 3 views as input. What’s more, these artifacts were related to the observed angle of input views, which induced the directional distortion denoted by the frequency spectrum and background interference at the lateral projection. As the number of views increased, the occurrence of artifacts gradually reduced (less in 7 views and diminished in 19 views). However, the deconvolution results still exhibited a loss of structural details and unnatural contrast characterized by the high-frequency noise in spectrum and low SSIM (~0.8). As for LF-INR, leveraging its continuous representation derived from model bias effectively mitigated undesirable hallucinations caused by a low angular sampling rate, and produced a higher structural similarity compared to deconvolution results (Fig. 3c). Furthermore, it is noteworthy that the structural fidelity of the LF-INR reconstruction with "7-views" surpassed that of the deconvolution with "19-views", highlighting the robustness of LF-INR in preserving structural details despite variations in the quantity of input views.

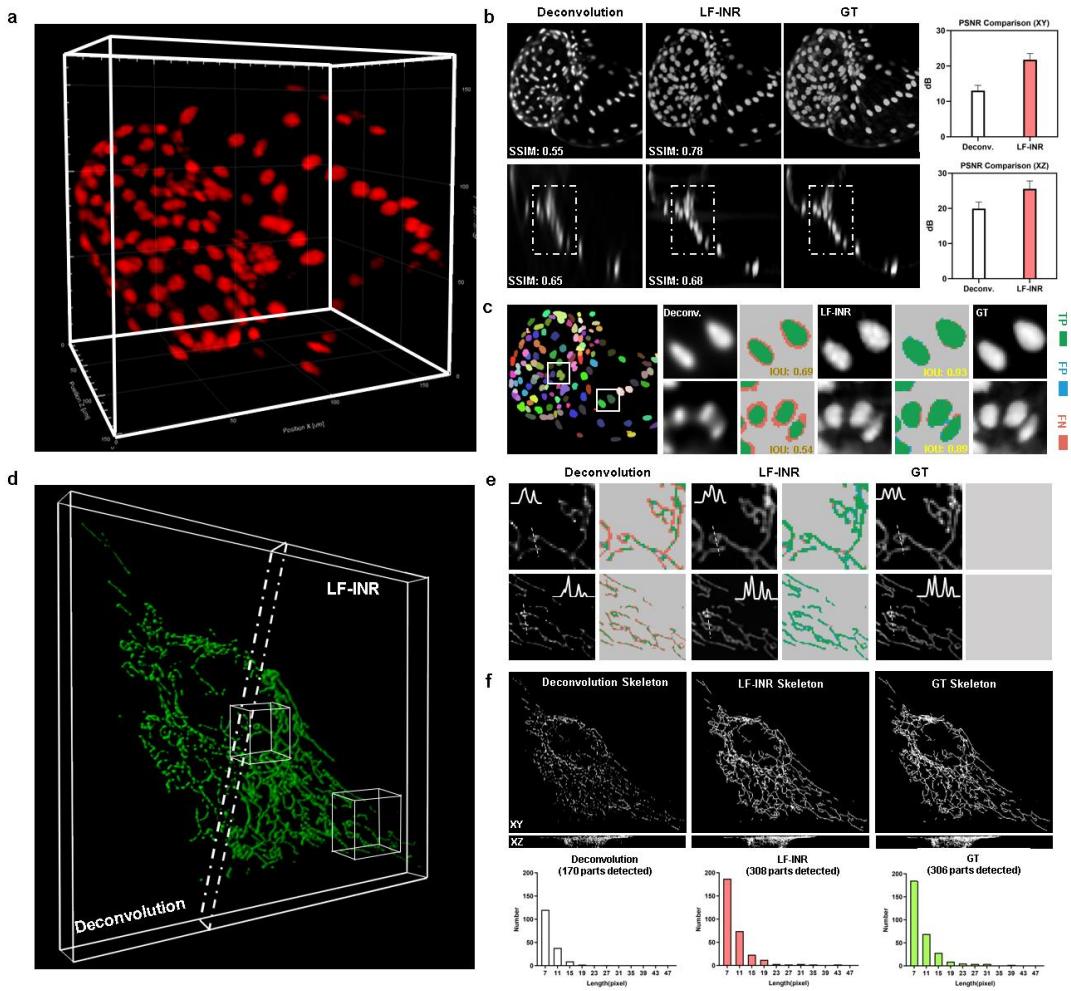


Fig.4 LF-INR enables high-fidelity reconstruction across diverse biological sample. a, LF-INR prediction from simulated FLFM projection of cardiomyocyte nuclei. **b,** Lateral (upper) and axial (bottom) projection of reconstructed volume in **(a)** by deconvolution and LF-INR. **c,** The segmentation results of reconstructed volume in **(a). d,** LF-INR prediction from simulated FLFM projection of mitochondria. **e,** Visual comparison of 3D reconstruction on ROIs marked in **(d)** by deconvolution and LF-INR. Corresponding color-encoded error maps were calculated with ground truth data. The intensity plots denoted LF-INR method outperforms deconvolution in structure details. **f,** *Upper: The* skeleton extraction results of deconvolution results, LF-INR prediction and ground truth. *Bottom*: The statistical distribution of detected mitochondria.

High-fidelity reconstruction of LF-INR promotes the downstream analysis.
In the realm of biological imaging applications, the accuracy of downstream tasks is directly influenced by the results of the reconstruction algorithm. Therefore, in addition to image metrics comparison, we also conducted specific downstream tasks between the traditional deconvolution approach and LF-INR. Without loss of generality, we chose the mostly common signals: 'beads-like' structure (Cells, Fig. 4a) and 'line-like' structure (Mitochondria matrix, Fig. 4d) to evaluate the accuracy promotion brought by LF-INR's features.

In Fig. 4b, we compared the reconstruction quality of deconvolution and LF-INR from simulated 19 FLFM views (More details seen in Methods). Our method shows

higher structure similarity and PSNR (Peak Signal to Noise Ratio) than deconvolution approach. Further, the cell segmentation results based on pre-trained segmentation networks in Fig. 4c denote the higher IOU scores (Intersection over Union) of LF-INR. We attribute this improvement of structure fidelity to the powerful multi-view fitting and generation ability of LF-INR while the alternate iterative process of deconvolution will gradually influence the signal structure and distribution.

Besides the 'beads-like' structure, we also demonstrate the signal continuity outperformance of LF-INR on mitochondria matrix. As shown in Fig 4.d, LF-INR result preserve the signal continuity while deconvolution one shows over-sparsity and signal lossy. To further illustrate the difference, we also conducted the segmentation comparison of two results in the error map (Fig. 4e). Our method achieves lower FN and higher TP denoting the outperformance in signal continuity and fidelity. Furthermore, we applied skeleton extraction algorithm to visualize and quantify this continuity outperformance as shown in Fig. 4f. We use mitochondria segmentation algorithm (Methods) to obtain the statistical distribution of mitochondria length. The results show deconvolution reconstruction lose much weak signals compared with LF-INR one, especially for longer mitochondria.

Reconstruction of cellular dynamics offered by LF-INR

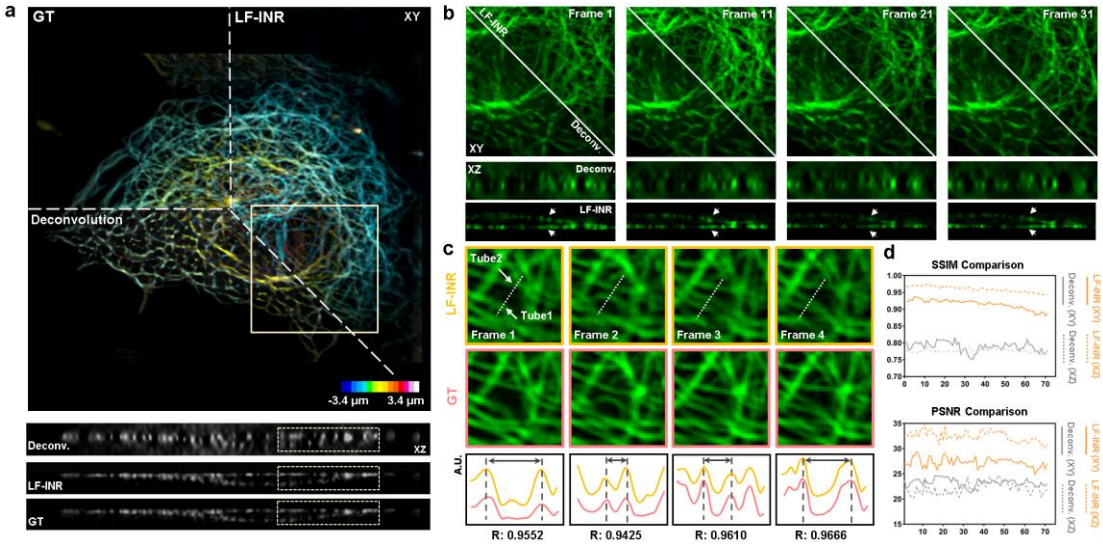


Fig.5 LF-INR enables capturing local dynamics assayed with simulated living cell. a, The depth-coded results of LF-INR prediction, deconvolution reconstruction and GT. **b,** Visual comparison between LF-INR and deconvolution results across different timepoint reconstructed from the marked regions in **(a)**. Frame interval was 2.3 s. **c,** The lateal projection of LF-INR results reconstructed from 4 consecutive frames. *Bottom:* The intensity plots denoting the changed distance between marked tubes. R is the correlation coefficient between LF-INR plot and GT plot. **d,** The statistical results of PSNR and SSIM calculated between GT and LF-INR predictions across 71 timepoints.

In living cell imaging, there existed biological dynamics across three-dimensional space, which revealed the modulation and organization of cellular systems. To evaluate

the performance of LF-INR on the dynamic signals, we conducted a simulated experiment of the reconstruction from motional microtubes. The 3D ground truth data were from real living cell imaging experiments (The time interval is 2.3 s), and the sequential reconstruction by LF-INR was based on transfer-learning strategy. As shown in Fig. 5a, LF-INR achieved higher signal continuity on the reconstructed tubes while deconvolution results suffered from the lossy of details at lateral projection and over-blurry at axial plane. Besides, this improved signal fidelity and axial resolution offered by LF-INR existed along sequential reconstruction as the white arrowheads denoted in Fig. 5b. To further quantify the continuity and quality of sequential LF-INR reconstruction by transfer-learning, two single tubes were selected (marked with white arrowheads in Fig. 5c) to measure the distance variation among four consecutive frames. With the improved structure fidelity, LF-INR achieved highly similar intensity distributions between selected tubes with GT ones (correlation coefficient is ~0.95), which promoted the precise location of selected tubes (the similar distance between two peaks in LF-INR plot and GT one).

Conclusion and Discussion

For multi-view 3D reconstruction in FLFM, how to decode the volumetric distribution with high structural fidelity under the input of limited && sparse angle captures is the key to interrogate the 3D dynamic morphologies. Classical analytical model-based approach can't solve this limitation due to the lack of any prior. While the supervised-learning based approaches bypass the physical model limitations via constructing a "2D-to-3D" mapping function under the guidance of volumetric data, but are laborious and costly for preparing massive labelled data, and suffer the weak generalization on new samples. Considering these, we proposed a novel self-supervised learning method, termed LF-INR, achieving high quality 3D reconstruction from FLFM captures with the internal continuity of INR model bias.

In comparison to the conventional deconvolution method based on an analytical model, LF-INR eliminated the axial artifacts existing in decoded dense signals yielded from multi-view reconstruction, and preserved high structural fidelity (SSIM: 0.9) and low errors (RMSE: 0.1). Notably, even in scenarios involving few-views reconstruction, LF-INR maintains high structural fidelity due to its inherent geometry consistency across predicted views. This view-synthesis ability of INR also mitigates the axial elongation induced by missing cone. With such improvements, our method outperforms in biological downstream analysis, like cell segmentation and mitochondria length statistics. Moreover, leveraging the correlation between FLFM and tomography, we extended this LF-INR framework to tomography reconstruction, demonstrating the robustness and universality of INR in solving the 3D reconstruction inverse problem (Supplementary Figure 7). We anticipate that this computational framework will facilitate the solution in other 3D imaging modalities.

Compared with vanilla NeRF, we modified the forward rendering model to match the wave-optics model in microscopy imaging. To mitigate the computational cost of voxel-by-voxel optimization, we implemented a "coarse-to-fine" sampling and decomposition strategy, enhancing the efficiency of 3D reconstruction. Additionally,

for time-variant signals, we employed a transfer-learning approach to expedite network convergence by leveraging temporal correlations, resulting in an acceleration of approximately tenfold. As a result, we demonstrate the outperformance of LF-INR compared with conventional approach. To enhance its utility, future efforts will concentrate on training acceleration[23], network generalization ability[22] and robust restoration under diverse degradation model [24, 25], in order to further extend its application in biological imaging.

Acknowledgments

This work was supported by the funding from the National Natural Science Foundation of China (T2225014, 21927802), National Key Research and Development Program of China (2022YFC3401102). The authors are grateful to Longbiao Chen and Yao Zhou for their helpful discussion.

Declaration of Interests

The authors declare no competing interests.

Methods

Simulation settings

In order to illustrating the features of LF-INR, we used simulated FLFM projections of various biological sample. To generate these multi-view projections, the 3D PSF of FLFM derived from wave-optics model simulation was convolved with high-resolution 3D stacks.

For generating FLFM projections of mitochondria, microtubes and ER, the magnification of objective was set as 100× and the focal length of Fourier lens was 400 mm. As for MLA (microlens array), the pitch was 3.25 mm and focal length was 120 mm. Under this settings, 7 views were generated.

For obtaining the projections of cardiomyocyte nuclei, we used 20×/0.5 objective and Fourier lens with 250 mm focal length. The focal length and pitch of MLA were 150 mm and 2 mm, which produced 19 views.

Network structure of LF-INR

The LF-INR network consists of coordinates factorization module, position encoding module and residual-MLP module as shown in Supplementary Figure 1.

Coordinates factorization module. The pre-defined grid containing the reconstructed signals is composed of $h \times w \times d$ coordinates, which means the subsequent MLP has to be queried with each voxel to predict corresponding intensity value. This significant computation demand quite influences the efficient implementation of 3D reconstruction[18]. Inspired by the factored decomposition of scene representation[26, 27], we combined the lateral plane representation (*xy*) and axial vector representation (z) to reduce the quantity of queried voxels. Specifically, in the training phase, the coordinates

in random-selected volume ($m \times n \times l$) at one iteration were split from (x,y,z) into (x,y) and (z), i.e., the computational cost was reduced from $m \times n \times l$ to $m \times n + l$. Different from total decomposition (namely, (x,y,z) to (x)-(y)-(z)), this unsymmetrical factored representation was derived from the redundancy or structural continuity among different Z planes which produce. Besides, anisotropic sampling and resolution decided by the imaging modality also denotes the difference of lateral and axial representation. We demonstrated this decomposition strategy outperforms original decomposition method[27] in Supplementary Figure 5.

Position encoding module. The coordinates' low-frequency variation among 3D space would degrade the rendering quality of structural details or color variation[18]. So, these decomposed coordinates also needed to be converted to high-frequency signal. In LF-INR, we adopt the conventional Fourier positional encoding:

$$\gamma(p_{xy}) = \{(sin2^0\pi p_{xy}, cos2^0\pi p_{xy}), (sin2^1\pi p_{xy}, cos2^1\pi p_{xy}), \ldots, (sin2^{L_{xy}-1}\pi p_{xy}, cos2^{L_{xy}-1}\pi p_{xy})\}$$

$$\gamma(p_z) = \{(sin2^0\pi p_z, cos2^0\pi p_z p_z), (sin2^1\pi p_z, cos2^1\pi p_z), \ldots, (sin2^{L_z-1}\pi p_z, cos2^{L_z-1}\pi p_z)\}$$

where P_{xy} is the lateral coordinate (x,y) while P_z is the axial coordinate (z). L_{xy} and L_z are lateral and axial encoding order, respectively.

Residual-MLP Module. In LF-INR, we adopted residual-MLP to learn the latent function mapping encoded coordinates to corresponding intensity. This module consisted of input layers, weights-sharing layers, feature fusion layers and output layers. The input layers consist of two branches which are used to extract primary feature of input encoded lateral and axial coordinates, respectively. Then, the weighs-sharing layers parallelly processed the extracted hidden feature from input layers. For different branches, they shared the same network weights within these layers based on the redundancy of grid representation. However, this accelerated factored representation would influence the view rendering quality compared with original 3D-grid optimization strategy [27]. Considering the performance degradation, we added two skip-connection branches to enhance the model ability as shown in Supplementary Figure 1. These skip-connections transmit the original input encoded information to the end of weights-sharing layers and the initial fusion layers to bypass non-linear transformation of intermediate layers and avoid information lossy brought by decomposition representation. Besides, for deep neuron network, these skip-connections also benefits the network convergence[28]. This reconstruction enhancement was demonstrated in Supplementary Figure 5.

Sequential 3D Reconstruction with weights transfer

The essence of INR is to use a neuron network to learn the mapping function from sampled coordinates to their corresponding values. For living biological imaging, the signal varies along 3D spatial coordinates and 1D time axis. Thus, only with the input spatial coordinates, INR can't represent the complex 4D function f (x,y,z,t). One plausible approach is to add the 't-variable' directly into INR optimization to learn the latent morphology pattern. However, this 4D representation increases the computational burden with extra sub-network and massive input coordinates. Besides, the irregular morphology and motion existing in 4D biological dynamics is hard to predict compared with regular linear motion or rotation[29]. Here, based on the high inter-

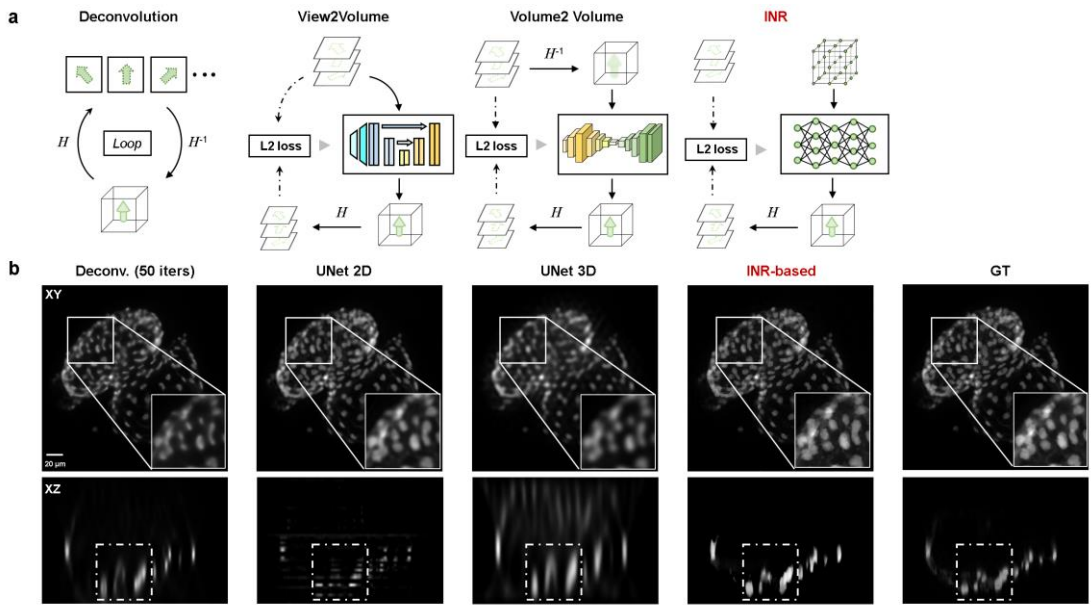
frame correlation, we adopt the transfer-learning strategy to accelerate the network convergence by sequentially fitting the multi-frame LF captures. During training phase, we first optimized the network to represent the first volume with the guidance of input T_1 views and got the trained parameters θ_1'. For the second volume, we initialized the network with θ_1', and the network was updated under the supervision of T_2 views. For remained volumes, according to consecutively transferring the previous optimized parameters, the network convergence time was dramatically reduced due to the high similarity between adjacent frames.

References

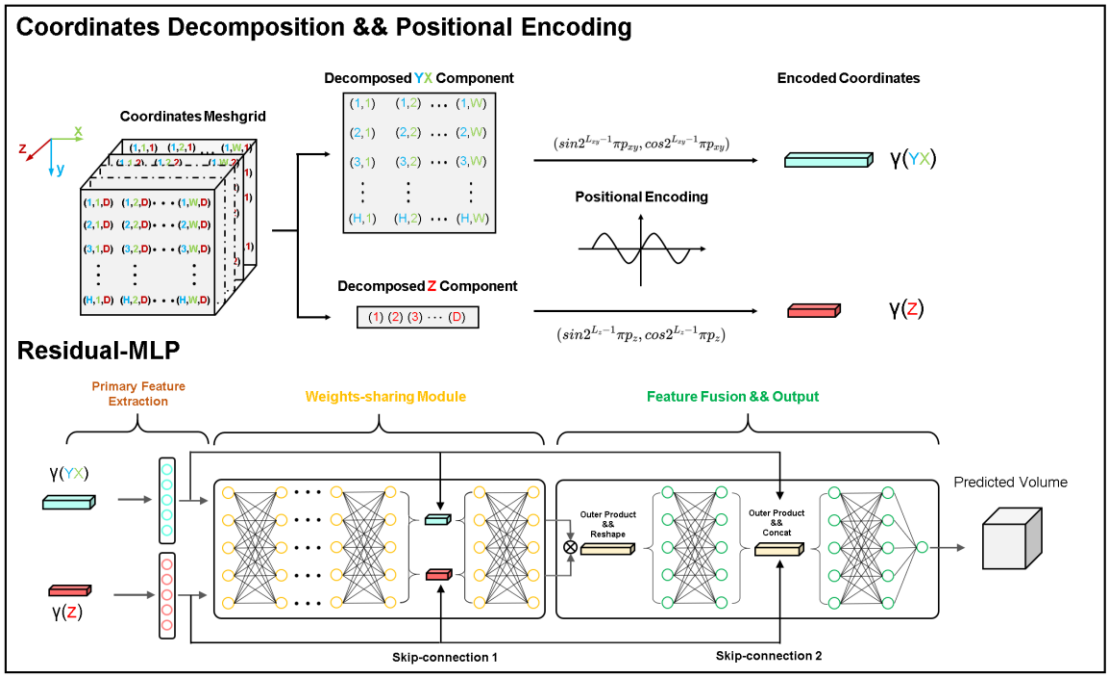
1. Voleti, V. et al. Real-time volumetric microscopy of in vivo dynamics and large-scale samples with SCAPE 2.0. *Nat. Methods* **16**, 1054-1062 (2019).
2. Mickoleit, M. et al. High-resolution reconstruction of the beating zebrafish heart. *Nat. Methods* **11**, 919-922 (2014).
3. Zhao, Y. et al. Isotropic super-resolution light-sheet microscopy of dynamic intracellular structures at subsecond timescales. *Nat. Methods* **19**, 359-369 (2022).
4. Qiao, C. et al. Rationalized deep learning super-resolution microscopy for sustained live imaging of rapid subcellular processes. *Nat. Biotechnol.* (2022).
5. Chen, B.-C. et al. Lattice light-sheet microscopy: Imaging molecules to embryos at high spatiotemporal resolution. *Science* **346**, 1257998 (2014).
6. Broxton, M. et al. Wave optics theory and 3-D deconvolution for the light field microscope. *Opt Express* **21**, 25418-25439 (2013).
7. Prevedel, R. et al. Simultaneous whole-animal 3D imaging of neuronal activity using light-field microscopy. *Nat. Methods* **11**, 727-730 (2014).
8. Hua, X., Liu, W. & Jia, S. High-resolution Fourier light-field microscopy for volumetric multi-color live-cell imaging. *Optica* **8**, 614-620 (2021).
9. Han, K. et al. 3D super-resolution live-cell imaging with radial symmetry and Fourier light-field microscopy. *Biomed. Opt. Express* **13**, 5574-5584 (2022).
10. Frikel, J. & Quinto, E.T. Characterization and reduction of artifacts in limited angle tomography. *Inverse Problems* **29**, 125007 (2013).
11. Kak, A.C. & Slaney, M. Principles of computerized tomographic imaging. (SIAM, 2001).
12. Wang, Z. et al. Real-time volumetric reconstruction of biological dynamics with light-field microscopy and deep learning. *Nat. Methods* **18**, 551-556 (2021).
13. Lanxin, Z. et al. Optics-aware super-resolution light-field microscopy for long-term volumetric imaging of dynamic intracellular processes at millisecond timescales. *bioRxiv*, 2023.2003.2015.532876 (2023).
14. Wagner, N. et al. Deep learning-enhanced light-field imaging with continuous validation. *Nat. Methods* **18**, 557-563 (2021).
15. Vizcaíno, J.P. et al. Learning to Reconstruct Confocal Microscopy Stacks From Single Light Field Images. *IEEE Transactions on Computational Imaging* **7**, 775-788 (2021).
16. Wang, F. et al. Far-field super-resolution ghost imaging with a deep neural network constraint. *Light: Science & Applications* **11**, 1 (2022).
17. Wang, F. et al. Phase imaging with an untrained neural network. *Light: Science &*

Applications **9**, 77 (2020).
18. Mildenhall, B. et al. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM* **65**, 99-106 (2021).
19. Yuan, Y.-J. et al. in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 18353-18364 (2022).
20. Li, Z. et al. in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 8456-8465 (2023).
21. Su, S.-Y., Yu, F., Zollhöfer, M. & Rhodin, H. A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. *Advances in Neural Information Processing Systems* **34**, 12278-12291 (2021).
22. Chen, A. et al. in Proceedings of the IEEE/CVF International Conference on Computer Vision 14124-14133 (2021).
23. Müller, T., Evans, A., Schied, C. & Keller, A. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)* **41**, 1-15 (2022).
24. Pearl, N., Treibitz, T. & Korman, S. in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 12672-12681 (2022).
25. Mildenhall, B., Hedman, P., Martin-Brualla, R., Srinivasan, P.P. & Barron, J.T. in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 16190-16199 (2022).
26. Cao, A. & Johnson, J. in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 130-141 (2023).
27. Liang, R., Sun, H. & Vijaykumar, N. Coordx: Accelerating implicit neural representation with a split mlp architecture. *arXiv preprint arXiv:2201.12425* (2022).
28. He, K., Zhang, X., Ren, S. & Sun, J. in Proceedings of the IEEE conference on computer vision and pattern recognition 770-778 (2016).
29. Pumarola, A., Corona, E., Pons-Moll, G. & Moreno-Noguer, F. in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 10318-10327 (2021).

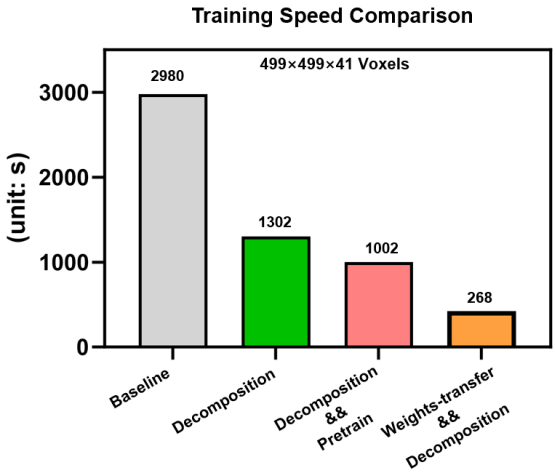
Supplementary Materials



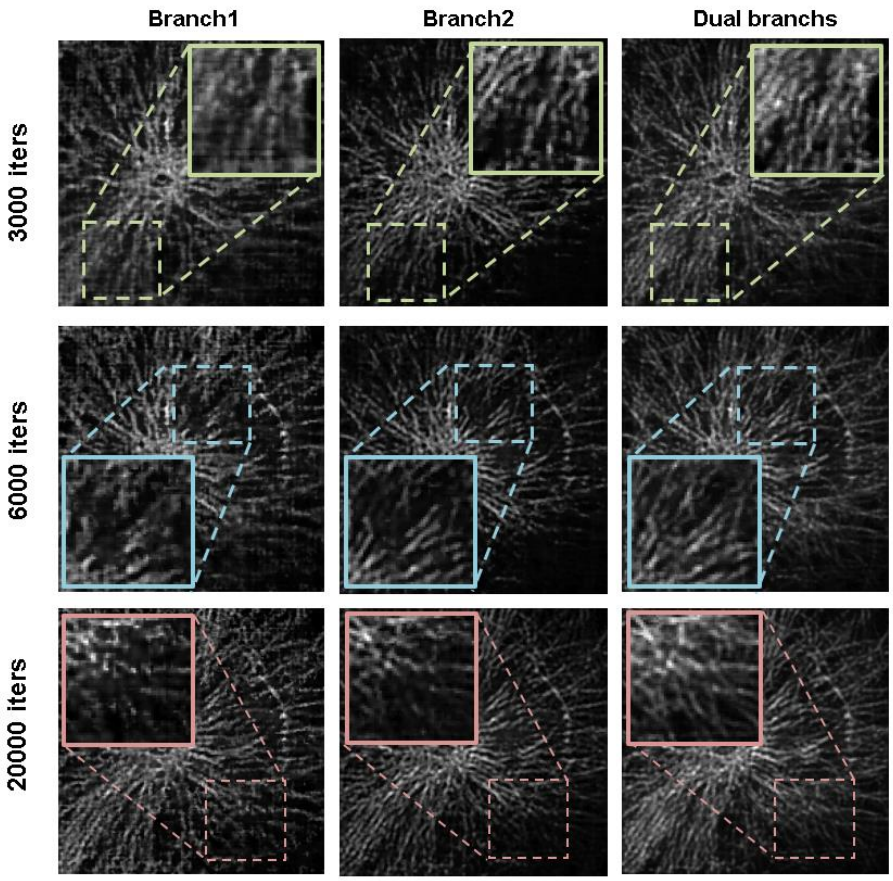
Supplementary Figure 1. The model comparison in self-supervised 3D reconstruction. **a,** Principle of 3D reconstruction approaches based on deconvolution and self-supervised learning. To our knowledge, most self-supervised 3D algorithms follow the paradigm of incorporation the physical forward model with a learnable neural network model. During training process, the loss function can be computed from input data and "re-projection" data (such as the convolution results of network prediction and PSF of pre-defined forward-model), which is called as "physical consistency". For example, the combination of VCD and FLFM PSF (*H matrix*), labeled as "View2Volume", can predict a 3D volume from input views under the constrains of the physical consistency. Further, by introducing an inverse matrix before network, another self-supervised algorithm can be obtained, which first converts input views into a "rough 3D volume" and refined this volume with a "Volume2Volume" network. **b,** Results of traditional deconvolution approach, "UNet 2D", "UNet 3D" and our proposed "INR-based" approach. The deconvolution results suffer structural distortion and axial elongation; "UNet 2D" improved the lateral structural fidelity but yielded axial discontinuity; "UNet 3D" preserve the axial continuity with volumetric prior but failed to refine this volume characterized by the axial elongation in the marked white rectangular area; Our approach performed high structural fidelity (seen in the XY projection) while mitigated the elongation and preserved high continuity in XZ plane.



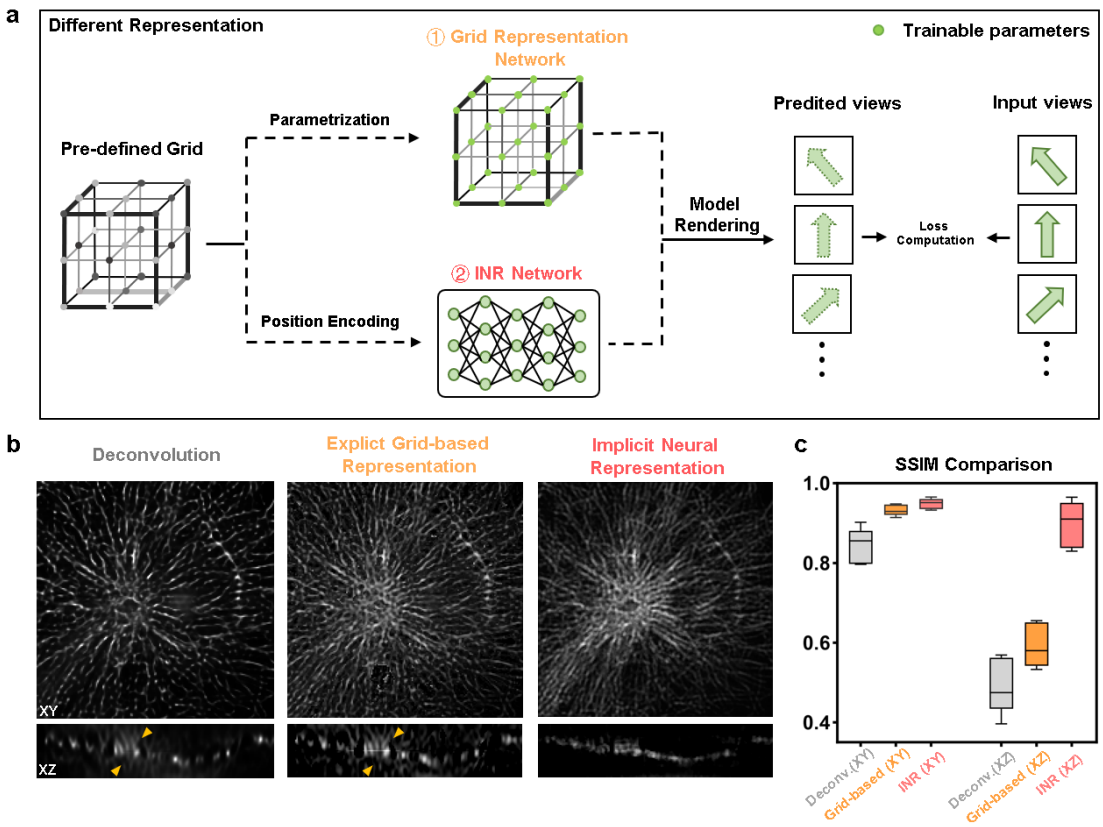
Supplementary Figure 2. The architecture of LF-INR network. Our proposed LF-INR is composed of "Coordinates Decomposition", "Positional Encoding" (above) and "Residual-MLP" (bottom). For the pre-defined coordinates, we first decomposed the (*x*,*y*,*z*) representation into two components: (*y*,*x*) and (*z*), inspired by the anisotropic sampling. Then, with the modulation brought by paired cos/sin function, the low-frequency information of coordinates changes across 3D volume can be converted to high-frequency one. When network training, these two encoded components were firstly sent to their own PFE (primary feature extraction) branch. With the sharing-weights layers, the primary feature was further processed yielding two intermediate outputs. Specially, before feature fusion, a skip-connection was built to convey the primary feature to intermediate outputs. In the fusion phase, an outer product was adopted to achieve dimension expansion for the latter transformation from 3d coordinates to voxel intensity. Similarly, the second long-skip connecting the primary feature with fusion features bypassed the no-linear operation of intermediate layers to avoid information lossy and promote network optimization. Finally, the output layer converted multi-channel features into voxel intensity. In our proposed LF-INR, the channel of feature extraction was set to 128 while the channel of output layer was 1 for gray channel prediction.



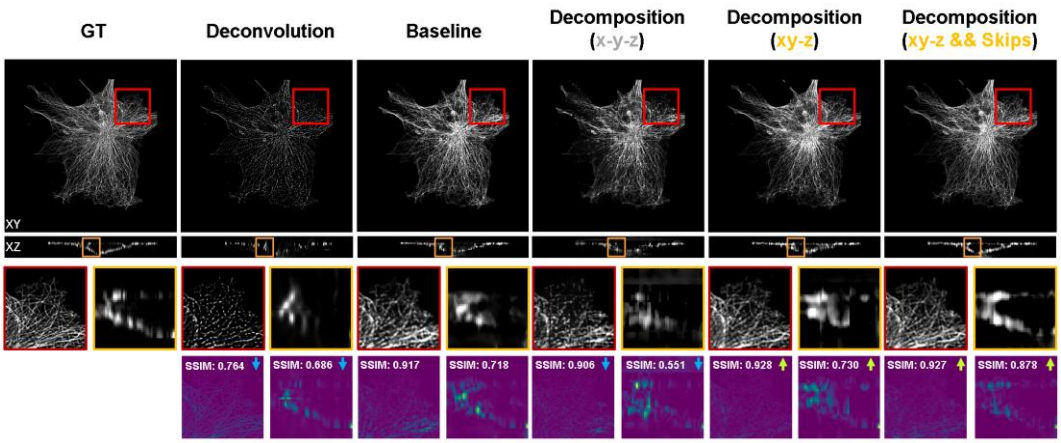
Supplementary Figure 3. **Speed comparison for network training.** With the joint optimization of training strategy and decomposed coordinates representation, our proposed network achieves about ~3 folds acceleration compared with conventional INR. Besides, the weight-transfer-based network optimization further reduce the convergence



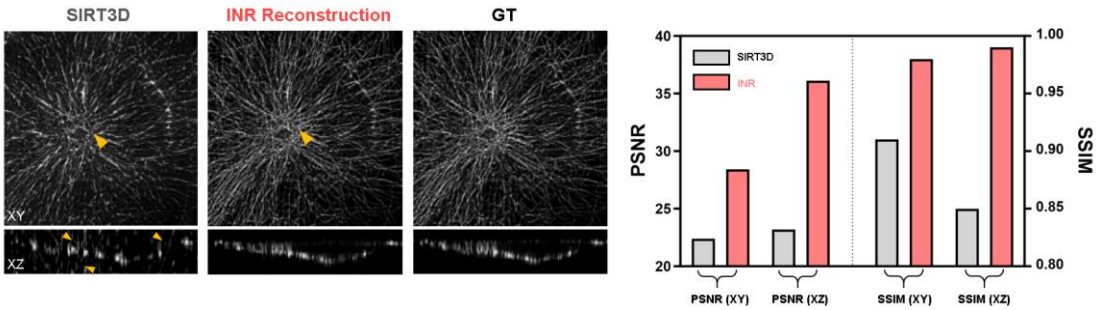
Supplementary Figure 4. **Hybrid rendering model promotes the reconstruction of structure details and network convergence.** Shown are maximum projection of 3D reconstruction via model branch 1 (wave-optics rendering model), branch2 (geometry rendering model) and dual branches (hybrid rendering model). In the same iteration, the results of "dual branches" outperformed single branch method in the structure details and fidelity as the ROIs shown.



Supplementary Figure 5. **The axial enhancement offered by the continuity of INR (implicit neural representation). a**, The framework of grid-based representation (explicit neural representation) and INR. Similar with deconvolution, grid-based representation initialized a volume whose voxels can be optimized via loss function. While INR adopted a MLP (Multi-Layer Perceptron) to learn a function which converted the vectors from encoded coordinates to corresponding voxel values. The intrinsic continuity of MLP acted a model prior to suppress the axial elongation and artifacts annotated with arrowheads. **c,** The SSIM statistical results of deconvolution, explicit grid-based representation and INR reconstruction. The mean SSIM are 0.8424/0.4870 (deconvolution at XY/XZ plane), 0.9330/0.5904 (explicit representation at XY/XZ plane) and 0.9488/0.9005 (implicit representation at XY/XZ plane).



Supplementary Figure 6. **Reconstruction comparison between different network design**. Results are shown for simulated microtube reconstruction under different algorithm settings, including deconvolution, baseline (conventional INR optimization strategy), "*x*-*y*-*z*" decomposition, "*xy*-*z*" decomposition and "*xy*-*z*" decomposition with long-skips. The bottom row indicates the error maps calculated with GT. Green arrowheads denote the structural similarity enhancement while blue ones denote quality degradation.



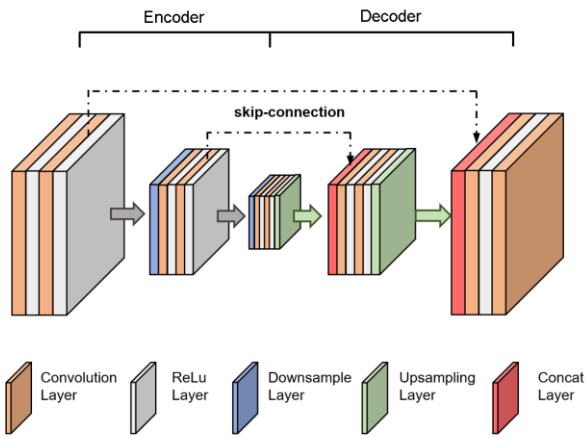
Supplementary Figure 7. **INR enables artifacts removal in tomographic 3D reconstruction**. In this study, the angle range of tomographic projection was set to the same value of FLFM settings and the number of views was 7. Different from FLFM, the simulated projection preserved the original resolution of ground truth. However, the missing cone problem and few views brought undesired high-frequency artifacts as denoted in SIRT3D results while INR reconstruction act finer structure details (yellow arrows at *XY*) and removed axial artifacts (yellow arrowheads at *XZ*). The plots of PSNR and SSIM (*right*) clarifies this quality improvement offered by INR.

Supplementary Note 1: Parallax extraction module:

In the hybrid forward rendering module, the second branch is applied to enhance the structural fidelity of projected views by geometry-based rendering model. Similar with the convolution process of FLFM PSF (point spread function), this rendering process can be described as the convolution between predicted volume and "parallax matrix". This matrix was generated by extracting the position of maximum value of FLFM PSF at different depth. Because of the view-correlated shift existing in PSF, this extracted position revealed the projective geometry relations among different views. Then, assigning a sliced PSF of center view at the focal plane to each position could yield a depth-independent PSF matrix representing parallax across different views.

Supplementary Note 2: View-enhancement network:

FLFM views suffered from the resolution degradation because of the blur brought by split aperture of each view, especially when meeting defocus signals. However, for in-focus signals (at the focal plane of objective), this degradation seemed to be weak and only related to the blur kernel at focal plane. Considering this, we built a view-enhancement network only based on input blurry views. First, we built the paired training images by simulating this degradation with in-focus signals: (1) Crop the in-focus ROI and apply 2D deconvolution algorithm to retrieve the degraded resolution (2) Blur the enhanced views ("HR") by various 2D kernels sliced along the depth of 3D PSF, which produced "LR" with different blur. (3) Crop the paired views to training dataset. Then, we trained a light-weight U-net (Supplementary Note Figure 2) with the paired data. With continuous optimization, the trained network can remove the defocus blurry and enhance the signal contrast.



Supplementary Note Figure 1. Network structure of view-enhancement network.
We adopted U-net composed of encoder and decoder to deblur the input views. Each encoder block contains 2 convolution layers with parameters n (channel number) and f (filter size) while each decoder block contains a bicubic upsampling layer and 2 convolution layers. The filter size and basic channel number in our experiment were set to 3 and 32, respectively.

Supplementary Note 3：Training details of LF-INR

LF-INR is designed to reconstruction 3D volume from captured 2D FLFM images in a self-supervised manner. Different from conventional CNN-based imaging process network, LF-INR is based on Multi-Layer Perceptron and converts the coordinates across 3D volume to corresponding grayscale value with the guidance of input views. The implementation of LF-INR contains three steps: 1. Fast training under the sparse sampling; 2. Refinement under original sampling rate; 3. Weights-transfer for sequential 3D reconstruction. The trained LF-INR acts a neural function enabling the transformation of "coordinates to intensity".

Due to the voxel-wise optimization strategy of INR-based network, the quantity of coordinates across 3D volume decided by spatial sampling rate influences the computational cost. Considering this, in the first training phase, we adopted a "down-sample version" of whole coordinates meshgrid as input to obtain a coarse 3D volume. Specifically, the coordinates across subsampling volume (×0.5) were sent to LF-INR to learn a "low-resolution" implicit representation. Besides, during forward projection, the 2D views and PSF function also need to be rescaled to match the sampling rate of the predicted volume. Due to continuity of MLP, the pre-trained volume can be interpolation by network prediction smoothly, which reveals the promotion to the second stage training. In the second training phase, all coordinates of the volume were used to optimize the INR network. With numerous iterations, the structure details of predicted volume can be refined.

When reconstructing sequential 3D stacks, scene-wise optimization strategy of INR inevitably meets the increasement of computational cost decided by the number of captured FLFMs. Here, we adopt the weight-transfer strategy to accelerate the convergence speed of INR optimization. In particular, the optimized weights and bias are used to initialized the following LF-INR network. For example, with the two-phase training mentioned before, LF-INR acts the implicit volume representation of timepoint 1. For FLFM capture of timepoint 2, this initialized parameters brings about ~4 dB PSNR improvement when starting the optimization of "T_2: LF-INR".

The more detailed hyper parameters of LF-INR training are listed in **Supplementary Table 1**

Supplementary Table 1：Network hyperparameters of LF-INR

	Volume Size	Positional Encoding Order	Channel Number	Fully-connected Layer Number
Microtubules	499×499×41	8	128	8
Mitochondria Matrix	499×499×41	8	128	8
Cardiomyocyte Nuclei	307×307×51	8	128	8