High-fidelity, generalizable light-field reconstruction of biological dynamics with physics-informed meta neural representation

Chengqiang Yi[1,7], Yifan Ma[1,7], Minglu Sun[1,7], Xinyue Yuan[1], Lanxin Zhu[1], Jiahao Sun[1], Shangbang Gao[3], Meng Zhang[4], Yuhui Zhang[4], Zhaoqiang Wang[5], Hsiai Tzung[5,6], 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*

[7]*These authors contribute equally to this work*

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

Abstract:

Light field microscopy (LFM) offers rapid volumetric imaging of dynamic biological processes via 2D light field snapshots. However, LFM's 2D-to-3D inversion with limited spatial-angular tomography leads to artifacts and poor axial resolution. Here, we introduce light-field meta neural representation (LFMNR), a new LFM reconstruction paradigm based on physics-informed implicit neural representation and meta learning, to address these issues for LFM. Leveraging INR's continuity, LFMNR achieves self-supervised, artifact-free 3D reconstruction with enhanced structural fidelity (~2.2-fold improvement), spatial resolution (~4.4-fold enhancement) and data compression (~10-fold), when compared to classical model-based light-field deconvolution. In addition, our meta-learning and progressive sampling strategies also mitigate INR's intrinsic limitations in weak generalization and low representation speed scene by scene, thereby resulting in rapid representation (~100-fold acceleration) of hundreds of consecutive volumes capturing sustained biological dynamics in three dimensions. LFMNR demonstrates superior performance across diverse biological samples without any prior spatial structure knowledge. We showcase LFMNR's

capability in observing cell apoptosis in several hours and capturing instantaneous organelle interactions at millisecond timescale. LFMNR approach readily delivers high-fidelity, high-speed 3D imaging with vast potential applications in biological research and provides a paradigm shift with extensive value for computational imaging reconstruction.

Introduction

Understanding the spatial dynamics and interactions within microenvironments necessitates high spatiotemporal resolution across three dimensions. This imperative drives the advancement of volumetric imaging technologies to probe rapid biological processes, such as cardiac hemodynamics[1, 2] and intracellular interactions[3-5]. While scanning-based microscopy compromises volumetric imaging speed and can induce additional phototoxicity due to its sequential scanning approach, light field microscopy[6] (LFM) enables one-shot volumetric imaging by capturing all fluorescence signals from an excited volume in a single 2D image through multi-view captures. Though LFM is ideal for imaging dynamic biological processes[7], the non-uniform spatial resolution resulting from frequency aliasing limits its broader application. Fourier light field microscopy[8-10] employs a microlens array at the Fourier plane to mitigate spatial sampling aliasing in standard LFM and better leverage the optical aperture, thereby achieving improved spatial resolution with fewer views. Since these methods commonly employ Richard-Lucy (RL) deconvolution for iterative 3D volume reconstruction using the point spread function, challenges arise under the lower angular sampling rate and limited angle-range of FLFM. This alternate projection algorithm, applied in a tomography reconstruction manner, may introduce artifacts and impact structure fidelity, especially when the signals are dense[11, 12].

To address limitations in multi-view 3D reconstruction, data-driven algorithms are proposed, leveraging deep learning models to convert 2D LF captures into high-resolution 3D stacks with improved reconstruction quality[13-17]. However, end-to-end supervised approaches require large-scale volumetric data and well-aligned "2D-3D" training pairs, which are not readily accessible and prone to introducing artifacts due to mismatch between training and experimental data[18]. Self-supervised learning-based algorithms, which do not require volumetric data, offer a promising alternative for advanced light-field reconstruction. Current self-supervised methods in microscopy tasks often integrate physical priors into models through well-designed loss functions derived from analytical models to ensure physical consistency[19-21]. However, building self-supervised models for 3D reconstruction encounters challenges primarily due to the dimensionality reduction, which adds complexity to finding viable solutions within physical consistency constraints. This complexity often leads to significant artifacts, hallucinations, and discontinuities in the reconstructed output (Supplementary Figure 1, Supplementary Note 1).

Recently, Neural Radiance Field[22] (NeRF) has been known for its ability to generate novel views with geometry consistency, and applied in various vision tasks[23-26]. NeRF employs an MLP (Multi-Layer Perceptron) to represent 3D scenes from 2D multi-view input, known as implicit neural representation (INR), capturing the latent

continuous function of signal distribution. While NeRF's volumetric representation correlates with the 3D reconstruction inverse problem in FLFM, differences in forward models between macrophotography and diffraction theory in microscopy imaging pose big challenges for realization of INR-based light-field reconstruction. Moreover, NeRF's grid-based training strategy is highly computationally intensive, limiting its further application in processing dynamic data.

Here, we introduce a physically informed, meta-learning based self-supervised approach, termed light-field meta neural representation (LFMNR), for efficient light-field 3D reconstruction with high fidelity. By integrating light field forward model into implicit neural representation for the first time, light-field INR (LFINR) achieves 3D reconstruction through self-supervised learning from 2D views. Leveraging the continuity of the INR function and the geometry consistency of synthesized views, LFINR effectively reduces artifacts and axial elongation even with extremely-limited views input of just 3, far surpassing classical model based deconvolution approaches. In addition, we employ a progressive sampling strategy with structural error maps to significantly reduce computational time compared to conventional INR methods. Furthermore, based on the intrinsic continuity of dynamic LF time-sequences, we utilize meta-learning to transfer optimized weights to adjacent time points, achieving accelerated (~100-fold faster) light-field meta neural representation (LFMNR) for high-throughput reconstruction of sustained light-field videos. Several experiments have validated LFMNR's minimized artifacts and improved reconstruction fidelity across diverse biological samples. It achieves approximately 2.2-fold improvement in structural fidelity and a 4.4-fold enhancement in resolution compared to conventional LFM deconvolution. Long-term cell apoptosis and several intracellular dynamic events such as mitochondrial fission and mitochondria-lysosome interaction cell apoptosis, have been identified and quantitatively analyzed to demonstrate the capability of LFMNR.

Results

The principle of LFINR

The framework of LFINR includes 3 steps: spatial coordinates sampling, scene-representation network constructing, and physics-model consistency evaluation, as illustrated in Fig. 1. Unlike conventional convolution neural networks (CNN) in image enhancement or reconstruction, LFINR requires spatial coordinates $\{x_i, y_i, z_i\}_{i=1}^{N}$ as network input, which means a 3D spatial grid covering the reconstructed signals needed to be defined before network inference ("*Objective space*" in Fig. 1a). However, due to the discrete grid form of image data, the quantity of spatial coordinates will increase multiplicatively with the enlargement of image size, which burden the computation cost and reduce network convergence speed. Here, we introduced a progressive sampling strategy, which first used down-sampled mesh grid to pre-train network and next fed all grids for details fitting, to make network fast convergence. Initially, a sparse-sampled mesh grid was employed to pre-train the network, providing a preliminary approximation of the overall distribution ("*i. Sparse sampling*" in Fig. 1b). Subsequently, during the fine-tuning phase, the full-size grids were firstly fed into the

network for details recovery ("*ii. Full sampling*" in Fig. 1b). Then, during the later stages of training, a focal sampling approach was implemented to prioritize the training on regions exhibiting higher distortion, in which, we leveraged local structural similarity to compute the error map ("*iii. Focal sampling*" in Fig. 1b, Methods). With this spatially aware optimization strategy, LFINR could achieve ~2-fold acceleration and outperform in weak-signal reconstruction compared with random-sampling based approach[22] (Supplementary Figure 3).

After coordinates sampling, a fully-connected network ("MLP network" in Fig. 1b) was constructed to serve as the underlying continuous volumetric scene function. By querying the 3D spatial coordinates, the network could recover the volumetric intensity information.

Finally, the produced volume was utilized to generate projections with forward rendering model for physical consistency evaluation. However, during this process, the difference between forward rendering model and real-world model degraded the 3D reconstruction. For example, the photography model utilized in conventional NeRF[22] relies on ray-casting model to form the volumetric density and color of opaque 3D scene. While FLFM imaging model is based on wave-optics theory and aims to accurately describe the intensity distribution of semi-/transparent sample. The ignoration of diffraction effect in wave-optics model will blur the represented 3D volume with increased errors (Supplementary Figure 4).

Taking this into account, we developed a hybrid physics model following the principle of wave-optics model and tomographic projection geometry to obtain dual-channel forward projections in rendering process and compute the hybrid loss function for evaluating physics consistency (Fig1b, Methods). For the wave-optics rendering branch, the FLFM Point Spread Function (PSF) was employed to convolve with the predicted volume to generate wave-optics rendering results, which simulated the light-field propagation determined by the hardware settings ("*Light-field propagation*", Fig. 1a). For the geometry rendering branch, we derived the projection matrix through parallax extraction and obtained the geometry rendering results (Methods). During the model optimization phase, we adopted weighted hybrid function based on the dual rendering results. The first term of loss function was pixel-wise errors (L_2 loss) between the extracted views and wave-optics rendering results. The second one was structure discrepancy to capture the skeleton information of input views. Before this loss computation, we employed a self-supervised view-enhancement network, termed as VE module in Fig. 1b, to remove the depth-correlated blur in extracted LF views, which matched the geometry rendering model (Supplementary Note 2). Then, the structure discrepancy could be derived from Structure Similarity Index Measurement (SSIM) between enhanced views and geometry rendering results. By continuously minimizing the hybrid loss, LFINR gradually learned to produce the 3D volume only with input coordinates, in a form of scene representation. Our hybrid physics projection model, compared to utilizing only one rendering strategy, exhibited accelerated network convergence, and maintained high structural fidelity (Supplementary Figure 5). Further details about the network implementations can be found in the Methods section.

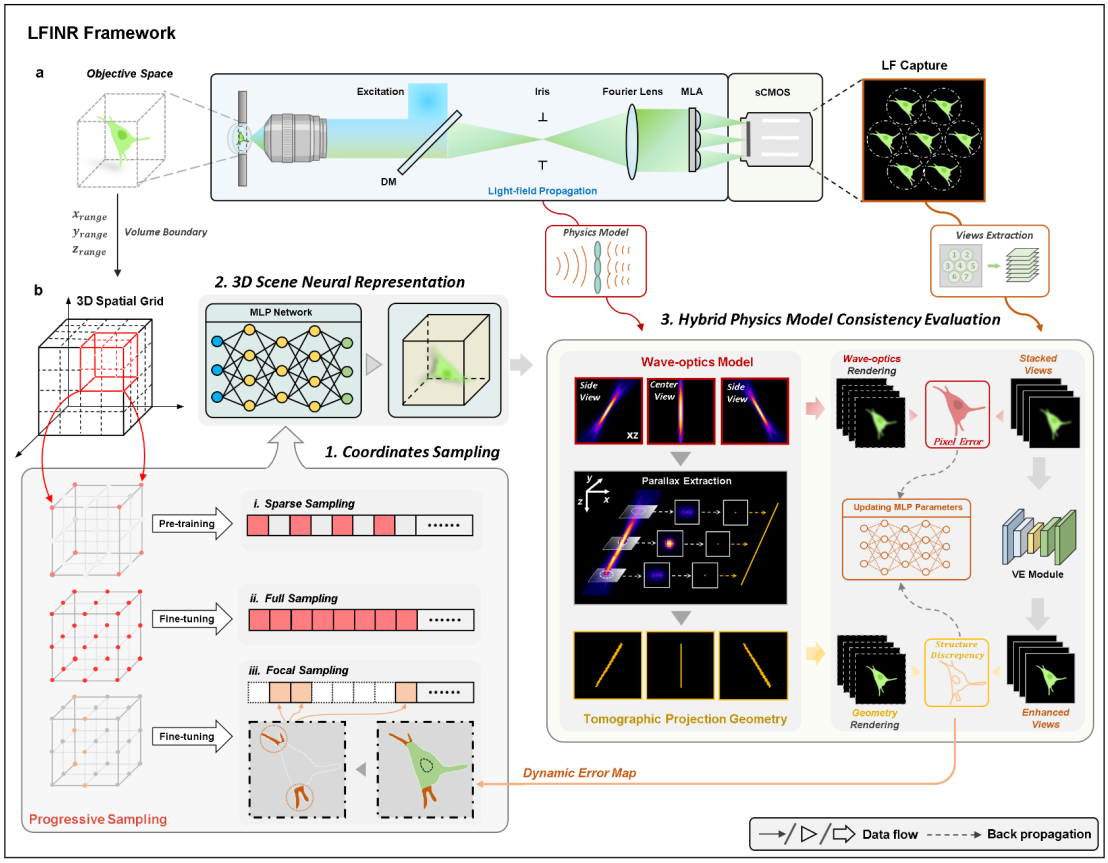

Fig1. The overview of LFINR framework. a, The schematic diagram of the Fourier light-field microscopy. **b.** Workflow of LFINR, which contains 3 steps: Sampling coordinates from 3D spatial grid with progressive strategy; Constructing MLP (Multi-Layer Perceptron) based scene-representation network; Computing the hybrid physics model consistency. With constant network optimization in a self-supervised manner, LFINR could learn to represent the 3D scene within the MLP, referred to implicit neural representation. During the optimization process, progressive sampling strategy was adopted to accelerate network convergence under the guidance of structural error map. The hybrid physics model contains wave-optics model for evaluating pixel errors between the extracted views and wave-optics rendering results, and tomographic projection geometry for evaluating structure discrepancy between geometry rendering results and enhanced views obtained by VE (view-enhancement) module.

LFINR enables high-fidelity 3D reconstruction by its continuous representation. For the inverse problem of 3D reconstruction from encoded multi-view light-field images, the quality of reconstructed results is highly depended on the performance of decoded algorithm. The 3D supervised DL algorithms[13, 15, 17] solve such problem in an unexplainable way by building CNN network mapping from 2D capture to 3D volume, which suffers the generalization errors due to the data bias[18]. The classical model based method, such as RL-deconvolution[8], is still the most widely-used approach especially when the volumetric data guidance is not available. However, the inherent constraints within inverse problems, characterized by limited angle range and sparse sampling of captured views[11], often lead to poor reconstruction quality, particularly evident along the axial dimension and for densely packed signals.

We compare the reconstruction between FLFM deconvolution and our LFINR on different biological samples as shown in Fig 2. In Fig. 2a, LFINR preserved obviously more structure details and higher contrast compared with deconvolution approach when decoding the 3D microtubule signals in an *U2OS* cells (labeled with 3mEmerald-Ensconsin) from 7 simulated FLFM views (Methods). The deconvolution approach exhibited noticeable blurs in lateral/axial projection and significant view-related stripe artifacts, which have been validated by the frequency analysis and high structure errors. Such deficiencies came from the depth-wise blurry modulation and signal cross-talk in Fourier light-field imaging process. For suppressing the blur effect, we first used the self-supervised VE network to remove the resolution degradation in FLFM views (Methods). The blur occurred in traditional deconvolution results decreased remarkably but the stripe artifacts still existed, as shown in "*View-enhanced Deconv*.", Fig 2b. In sharp contrast, LFINR produced artifacts-free 3D reconstruction with higher SSIM of ~0.96 and lower NRMSE (Normalize Root Mean Square Error) of ~0.058, and recovered much more high-frequency details validated through the spectrum analysis (insets in Fig. 2b). These notable enhancements in quality stemmed from the intrinsic continuity of implicit representation and physics-based hybrid rendering model in LFINR (Supplementary Figure 6).

In addition to the improvements in structural fidelity and spatial resolution, LFINR, with its ability for arbitrary view synthesis facilitated by the continuity of representation, also demonstrate significant superiority in 3D reconstruction of ER (labeled with EGFP-Sec61β) with extremely low number of light-field views, such as just 3, achieving ~2.3-fold higher SSIM, ~1.3-fold improvement in xy planes and ~1.76-fold improvement in xz planes (Fig. 2c, Supplementary Figure 7). The deconvolution-based results exhibited severe striped artifacts in axial plane (white arrowheads), when only 3 views were used. Such undesirable hallucinations across different depth further deteriorated the lateral structural fidelity denoted by the lower SSIM (Fig. 2c, d). Besides the evaluation in space domain, we also conducted geometry consistency comparison among different results (Fig 2e, Method, Supplementary Figure 8). The EPIs (epi-polar plane images)[27] used here described the distribution of both angle and space, and the slopes of lines in EPIs denoted the parallax across different views. When angular sampling was reduced, deconvolution-based methods exhibited artificial stripes in EPIs. In contrast, LFINR's outputs maintained a seamless transition across various views, as evidenced by the consistent alignment of signals and a higher SSIM of 0.908. This indicates LFINR's capability to generate arbitrary views with strong geometric consistency, even when operating under constraints of limited supervision

We noted that, with the image quality enhancement, LFINR also achieves higher accuracy in downstream analysis (Fig. 2f, g). The segmentation results of cardiomyocyte nuclei (labeled with *Tg(myl7:nls-gfp)*) showed higher intersection over union (IOU) scores of LFINR, while deconvolution ones showed obvious errors in the boundary of single cell (red area in "*Deconv*" results, Fig. 2g). Additionally, with the alternative projection process in RL-deconvolution, structural errors tended to gradually accumulate, as evidenced by the over-refined details. (Supplementary Figure 9).

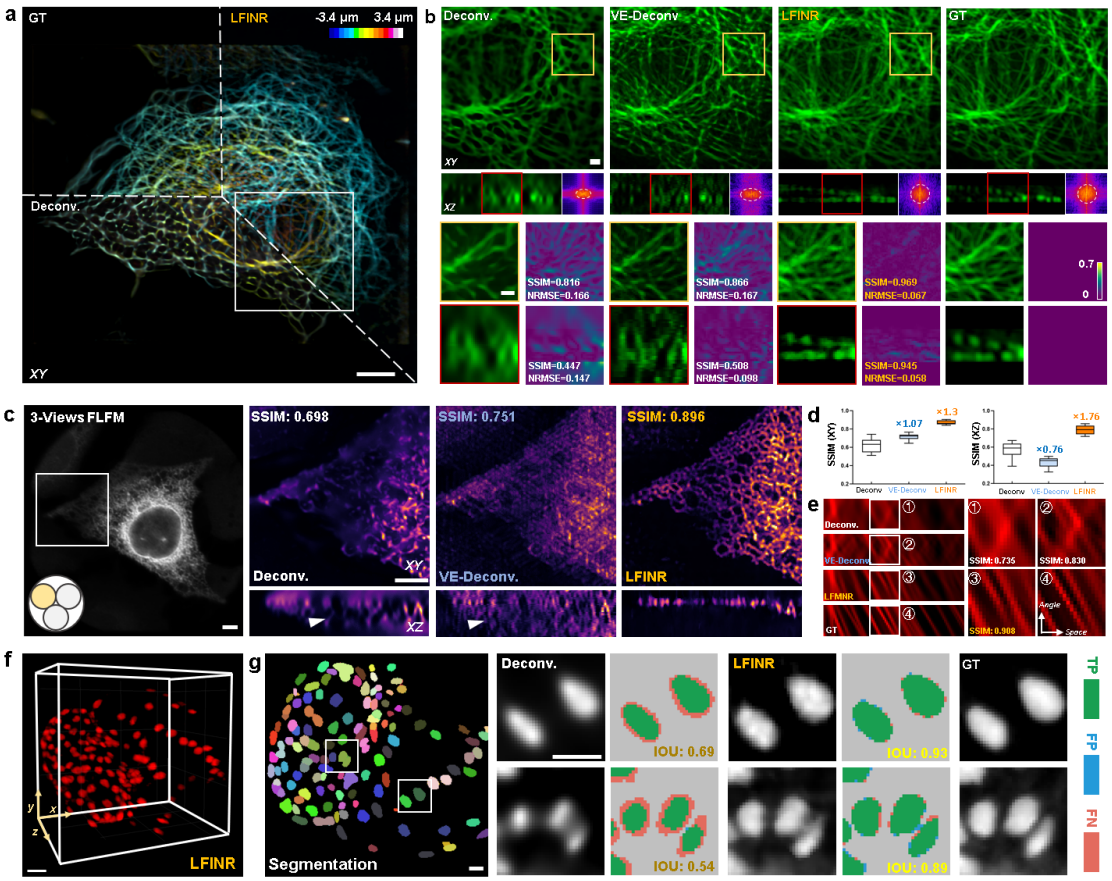

Fig.2 Light-field reconstruction with improve quality by LFINR. a, The depth-encoded projections of microtubules by LFINR reconstruction, FLFM deconvolution and ground truth (GT). The number of FLFM views used was 7. Scale bar: 50 μm. **b,** magnified horizontal and vertical sectional views of white rectangular boxes marked in **a**, denoting the higher structure fidelity of LFINR results compared to deconvolution and VE-Deconv (view-enhanced deconvolution) ones. Scale bar: 2 μm. **c,** The comparison between deconvolution, VE-Deconv and LFINR results reconstructed from three FLFM views of ER (endoplasmic reticulum). Scale bar: 5 μm. **d,** Structure similarity comparison of three approaches. The numbers above the "VE-Deconv" and "LFINR" boxes represent the structure similarity enhancement ratio compared to the deconvolution approach. **e,** Angular consistency evaluation through computed EPIs (epi-polar plane image) comparison of different approaches. Top to bottom: deconvolution, VE-Deconv, LFINR and GT. **f,** LFINR reconstruction from 19 FLFM views of cardiomyocyte nuclei**.** Scale bar: 20 μm**. g,** The segmentation results of reconstructed volume in **(f)**, along with the comparison of IOU scores between deconvolution and LFINR. Scale bar:10 μm.TP: True Positive. FP: False Positive. FN: False Negative.

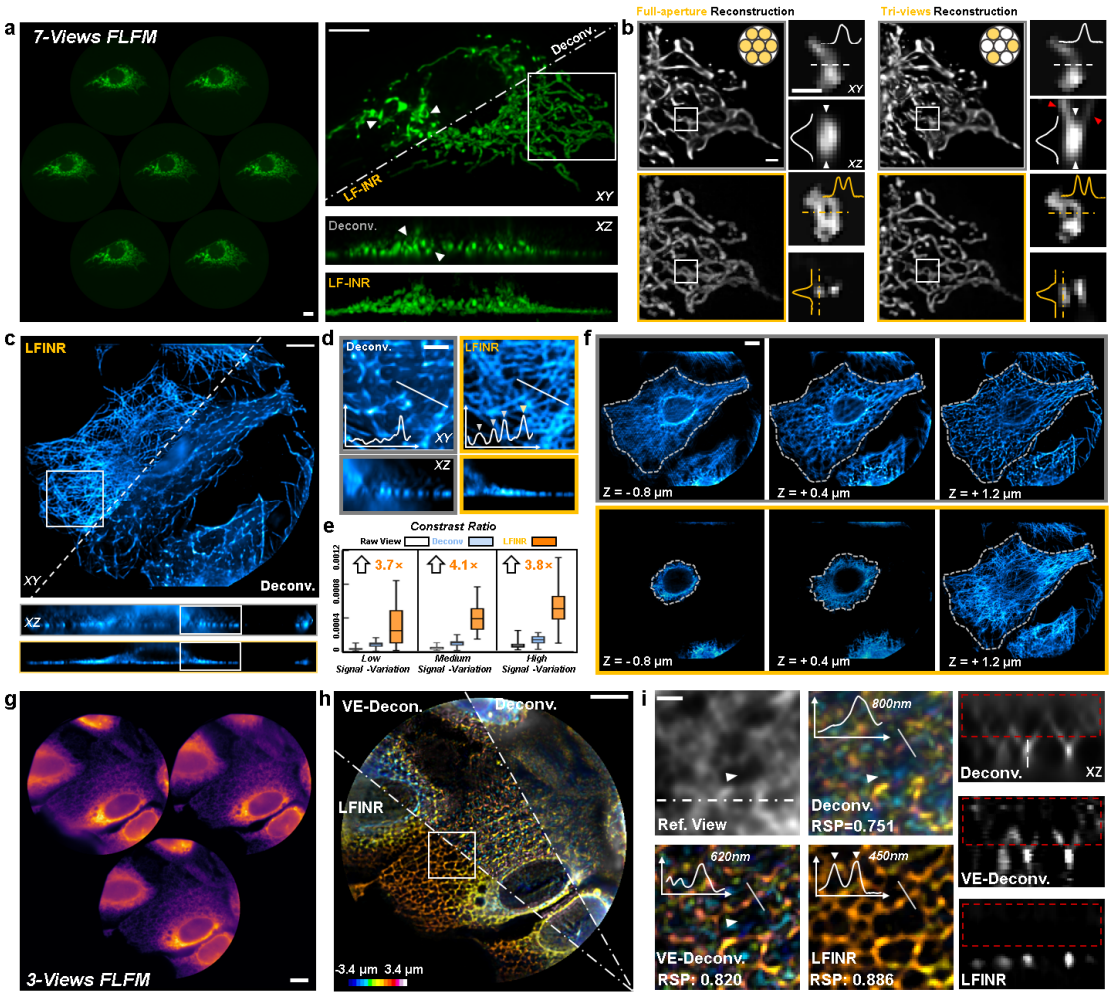

Fig.3 Comparative performance of deconvolution and LFINR reconstruction on diverse organelles. a, Captured views (7-view FLFM) of the mito Ma (tagged with Cox4-EGFP) in a whole live U2OS cell (left) and corresponding reconstruction results by deconvolution and LFINR (right). Scale bar: 10 μm. **b,** The comparison between deconvolution results (top row) and LFINR reconstruction (bottom row) in lateral and axial plane when the number of views becomes fewer. Full-aperture reconstruction referred to the reconstruction derived from 7 input views and tri-views reconstruction means using only 3 views (higher spatial resolution in each view) for the reconstruction. Insets denote the magnified view of *xy*/*xz* sections by each approach under each FLFM configuration. The linecuts through the resolved single mitochondria indicate the improved lateral and axial resolutions in LFINR reconstructions. Scale bar: 2 μm. **c,** Comparative results offered by deconvolution and LFINR along lateral and axial planes on microtubule (tagged with 3mEmerald-Ensconsin). Scale bar: 10 μm. **d,** The higher-magnification views of microtubule (marked by white rectangular regions in **c**). The plots denote the intensity profiles of the lines shown in the *xy* MIPs. Scale bar: 5 μm. **e,** The box plots show the statistical contrast ratio[28] of raw FLFM views, deconvolution results and LFINR results under various signal densities. The numbers denote the enhancement ratio compared with deconvolution approach. **f**, Comparison of the lateral planes at various axial position reconstructed by deconvolution (top row) and LFINR (bottom row). The dotted white line denoted the signal boundary at each depth, showing notably stronger optical sectioning capability by LFINR reconstruction. Scale bar: 10 μm. **g,** The captured 3 FLFM views of ER (tagged with EGFP-Sec61β) in a live U2OS cell. Scale bar: 10 μm**. g,** Depth-encoded *xy*

MIPS of 3D reconstructions by deconvolution, VE-Deconv and LFINR. Scale bar: 10 μm. **i**, The higher-magnification views of white rectangular regions in **g,** showing notably less hallucinations and axial aliasing, as indicated by the white arrows. The three insets shown in the right are *x-z* sectional planes along the white dash line in **i**. Scale bar: 2 μm.

Spatial- angular bandwidth enlargement offered by LFINR

The hardware design of Fourier light-field microscopy (FLFM) is an art of balance on angular and spatial information in a limited bandwidth determined by objective lens. In the image domain, such trade-off usually reflected as the form of spatial resolution degradation and axial artifacts. Here, in a complete self-supervised way eliminating concerns regarding generalization, LFINR showcased its strong reconstruction capability across diverse FLFM imaging configurations. The superior results were quantitatively verified through various metrics (Fig. 3).

As shown in Fig 3a, LFINR reconstruction revealed the detailed structure of mitochondria Ma (mitochondrial matrix) while deconvolution results exhibited blurry (denoted by white arrowheads) under 7 views FLFM settings. The line width measurement on resolved mitochondrial, in 'Full-aperture Reconstruction' of Fig 3b, denoted that our 7-view LFINR has surpassed the FLFM system's theoretical resolution limits, which shows ~1.68× and ~2.63× lateral and axial resolution enhancement compared to model-based deconvolution results. Besides, with the robustness in 3D reconstruction with few views, LFINR persevered the structural details while deconvolution yielded noticeable artifacts in axial plane (denoted by red arrowheads in '*Tri-views Reconstruction*', Fig 3b). Furthermore, with the increment of signals' complexity, the model-based deconvolution approach suffered the cross-talk from different slices. As depicted in Fig 3c, 3d, the microtube reconstructions by deconvolution suffered from the blur and axial elongation, thereby showing lower contrast and couldn't resolve the hollow structure in cell's axial slice. In contrast, LFINR recovered the sharp contrast of microtubules and enabled the 3D visualization of whole cell. The statistics results (Methods) in Fig 3e and different *xy*-planes in Fig 3f revealed the 3.7~4.1-folds contrast improvement under different signal's densities and the superior optical sectioning ability of LFINR.

Besides, LFINR significantly reduced the artifacts when reconstructing volume from merely three light-field views, under which though the views show higher spatial resolution (Fig 3g), excessive high-frequency hallucinations and cross-talk of depth information also arise (Fig 3h). These undesirable artifacts shown in deconvolution-based results (denoted by white arrowheads in Fig 3i) have led to inadequate RSP[29] metrics and low spatial resolution. In contrast, our LFINR removed the stripe artifacts brought by lower angular sampling rate and achieved enhanced fidelity indicated by higher RSP[29] values (0.886 vs 0.751, Fig 3i) and ~1.78-fold lateral resolution enhancement (450 nm vs 800 nm on dense Endoplasmic Reticulum signals , Fig 3i). High-quality 3D reconstruction with extremely few light-field views verified the powerful arbitrary view-synthesis ability of LFINR.

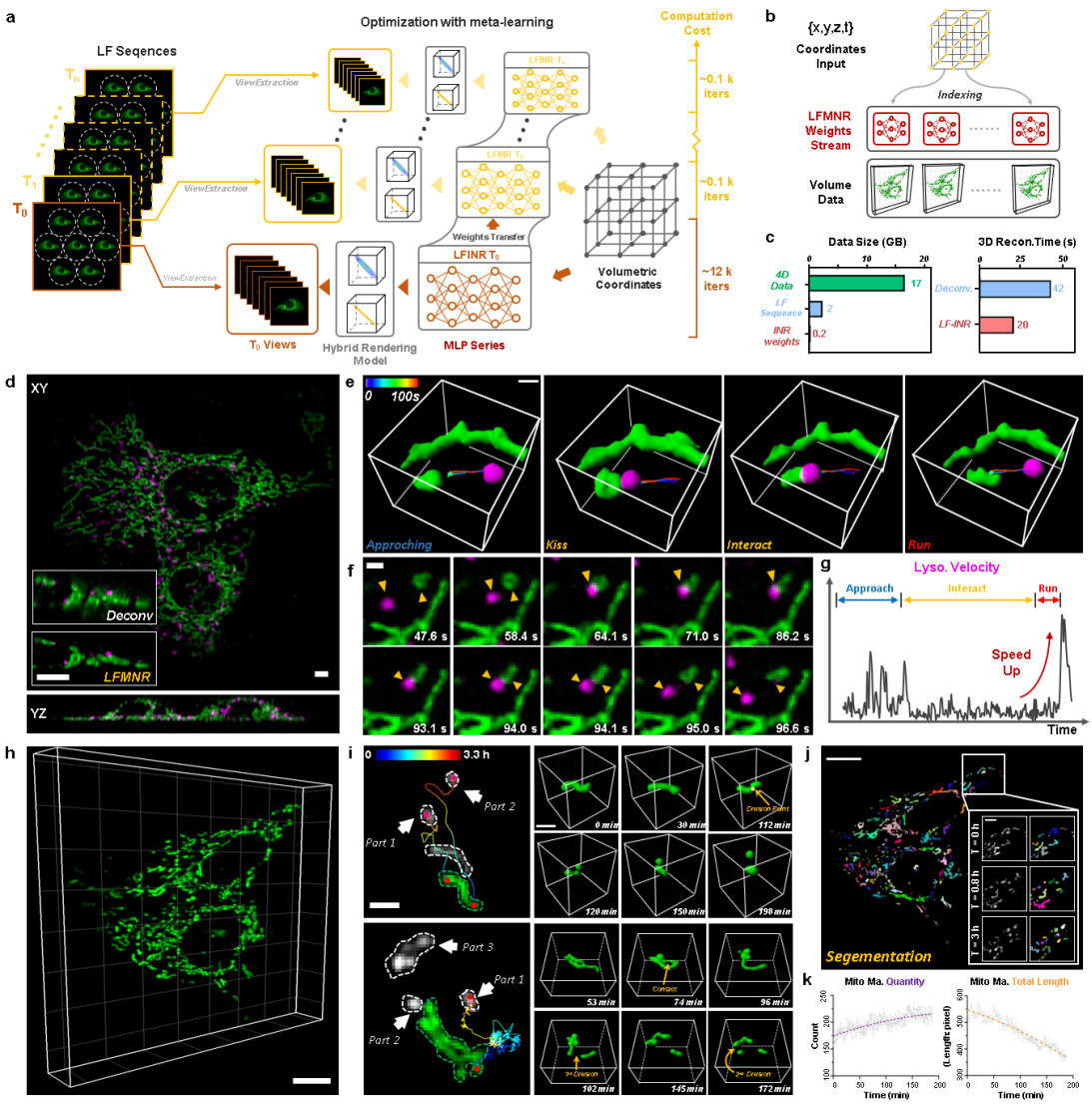

Fig.4 Visualizations and quantifications of intracellular dynamics in live cells by LFMNR. a, The scheme of light-field meta neural representation in solving multi-scenes generalization and accelerating optimization. **b,** The rendering process of 3D volume derived from LFMNR weights. **c,** *Left*: The data size of 4D data (3D videos), LF videos (raw LF sequences) and LFMNR weights (network parameters), denoting the strong data-compression ability of LFMNR. *Right*: The time-consumption comparison of deconvolution and LFMNR for 3D reconstruct of light-field vieos. **d,** Dual-color imaging of mitochondria (Green, tagged with Cox 4-EGFP) and lysosome (magenta, tagged with EGFP-rab7) in live U2OS cells. Insets show the comparison of axial section between deconvolution and LFINR. Scale bar: 5 μm. **e,** The volume rendering of the interaction process between mitochondria and lysosome occurring in 100 seconds. Scale bar: 1 μm. **f,** The lateral projection of dual-color LFMNR results illustrating the fast "kiss" (top panel) and "run" (bottom panel) process. Scale bar: 1 μm. **g,** The velocity changes of lysosome during the whole process visualized in **f**, quantifying the phenomenon of accelerated running away of lysosomes after detaching with mitochondria. **h,** LFINR reconstruction of the first frame of the long-term mitochondria evolution during cell mitosis. Scale bar, 50 μm. **i,** The dynamic process of mitochondria over an observation period of 3.3 hours. The orange-yellow arrowheads denote the spatial location of dynamics events. Scale bar: 2 μm. **j,** The segmentation result of LFMNR

reconstruction shown in (**h**). Scale bar: 20 μm. Insets show the fragmentation of mitochondria during observation. Scale bar: 5 μm. **k,** The statistical results of the total length and quantity of mitochondria during the whole observation based on the segmentation approach in (**j**).

Light-field meta neural representation (LFMNR) boosts the 4D visualization of intracellular dynamics.

Dynamic biological processes, such intracellular interactions, often occur across three-dimensional space for a long time. High-resolution extraction of the spatiotemporal patterns of these dynamics requires consecutive volumetric reconstruction of light field videos at high throughput, thereby posing big challenge to native LFINR which relies on the scene-wise representation. To make LFINR practical to continuous 4D visualization of biological dynamics, we proposed a weights-transfer strategy based on meta-leaning to enhance the efficiency of sequential 3D reconstruction by inheriting optimized parameters from prior implicit representation network, which was termed as LFMNR (Fig. 4a, Methods). Initially, the first frame (LFINR T_0) was served as the key-point to perform self-supervised 3D reconstruction with the progressive sampling strategy. Subsequently, the optimized parameters were employed to initialize the network of subsequent LF frame (LFINR T_1). Owing to the inherent motion continuity of biological dynamics, the convergence can be reached with merely 100 fine-tuning iterations with hybrid rendering model. As a result, an acceleration of ~100 folds was achieved as compared to the time consumption on native INR optimization, leading to an ~20 s quick reconstruction for each volume (Fig. 4c, Supplementary Figure 10). When sequential optimization finished, LFMNR could "memorize" the 4D information in a form of saved network weights. Users could input arbitrary 4D coordinates to index corresponding voxels (Fig. 4b) instead of storage of grid-based 4D images. With the unique feature of representation network and meta-learning strategy, our LFINR overcame the deficiency on reconstruction speed, meanwhile it achieved data compression (~10 folds reduction than the data size of LF views) and computation-cost reduction (~2 folds) compared with current deconvolution approach (Fig. 4c).

To demonstrate the utility of LFINR, we conducted dual-color live-cell imaging to capture the locomotion and interaction of mitochondria (tagged with Cox4-EGFP) and lysosome (EGFP-rab7) in three dimensions, as shown in Fig. 4d. We visualized the instantaneous 3D interactions between two types of organelles (Fig. 4e, f) at 10 Hz volume rate. We tracked the velocity of lysosome during the period of "approaching-contacting-separating", and identified the dramatic speed increase of lysosomes when they were running away (Fig. 4f, g). Besides the fast interaction between organelles, we also conducted long-term observation (~3.3 hours) of mitochondria evolution (Fig. 4 h, i). LFINR enabled capturing the multiple division events of mitochondria and completely tracking the daughter mitochondria. In addition, leveraging the high contrast of LFINR results, we segmented and identified each mitochondrion, revealing the smaller and rounder mitochondria resulted from extensive fragmentation of mitochondria shape (Fig. 4j). The statistical variation of mitochondria length and quantity in Fig. 4k also demonstrated the mitochondrial fragmentation process, indicating the potential apoptosis phase of the cells[30, 31].

Conclusion and Discussion

Achieving high-structural-fidelity 3D reconstruction in light-field microscopy, especially when the angular captures are limited and sparse, is essential for accurately capturing dynamic biological processes at high spatiotemporal resolution. Classical model-based approaches are inefficient in this regard due to the absence of prior knowledge. While supervised-learning methods offer a solution by constructing a "2D-to-3D" mapping function guided by volumetric data, they are laborious and costly in preparing labeled datasets and often suffer from poor generalization. Furthermore, the recovery of dense signals remains a challenge to all current light-field reconstruction approaches.

LFINR we propose in this work can achieve high-throughput and high-quality 3D reconstruction from a single 2D FLFM capture in a completely self-supervised manner, with the combination of inductive continuity bias of INR model and physics-based hybrid-rendering model.

Compared to conventional deconvolution approach, LFINR eliminates axial artifacts for dense signal reconstruction, maintaining high structural fidelity (Structural Similarity Index, SSIM: ~0.96) with low errors (Root Mean Square Error, RMSE: ~0.06). Notably, even in scenarios involving few-views reconstruction, LFINR maintains high structural fidelity due to its inherent geometric consistency across predicted views. INR's view-synthesis ability also mitigates axial elongation induced by missing cone problem, enabling artifact-free 3D reconstruction and spatial resolution enhancement by approximately ~4.4-fold in the imaging of diverse organelles. Furthermore, based on the correlation between FLFM and tomography, we extend the LFINR framework to tomography reconstruction, demonstrating its robustness and universality in solving a broad range of inversion problems in 3D image reconstruction (Supplementary Figure 11).

In addition to the improvement of reconstruction quality, to mitigate the computational cost in conventional uniform-random sampling strategy, we propose the progressive sampling with the aid of structural error map to accelerate convergence speed and prevent the signal loss. Further, to make LFINR practical in visualizing sustained biological dynamics, we develop a Meta-learning approach to expedite network convergence through temporal correlations, achieving ~two-order-of-magnitude significant acceleration with averaged inference time of merely 20 seconds for one volume.

In summary, our LFINR utilizes neural networks to smoothly represent high-dimensional scenes (3D volumes) with low-dimensional observations (2D FLFM views). This new "scene representation" paradigm solves the issue of inadequate quality in classical model-based light-field reconstruction approaches while completely eliminates the dependence on data priors. Meanwhile, it not only enhances the image quality for light-field 3D reconstruction but also exhibits capability in data compression. We envision that LFINR can be further optimized through the future efforts on training acceleration for real-time reconstruction[32], combination with high-quality optical setups for higher resolution beyond diffraction limit of the views[33], and robust

restoration under diverse degradation models[34, 35]. Besides the light-field microscopy, we anticipate that this self-supervised concept will also facilitate solutions in various other computational 3D bioimaging techniques, such as x-ray tomography[36, 37], cryo-electron tomography[38], Fourier ptychographic microscopy[39, 40], MRI[41, 42] etc.

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

FLFM optical setup

The FLFM system was built on a commercial inverted microscope (IX73, Olympus). To demonstrate the versatility of LFINR under various hardware settings, we conducted experiments with both 7-view and 3-view configurations.

For the 7-view setting, we utilized an objective lens (M60/NA1.5, Olympus UPLXAPO60XO) to collect fluorescent emission, which was controlled by an objective scanner. To prevent the overlap between captured views, an iris (SM1D12, Thorlabs) was positioned as the native image plane. The focal length of Fourier lens was set to 250 mm. A microlens array (*pitch* = 3.25 mm, *f-number* = 37, f_{ML} = 120 mm) was placed at the back foci plane of Fourier lens, producing 7 FLFM views. The whole schematic of the optical path is shown in Fig 1a.

For the 3-view setting, the objective (Olympus UPLSAPO100XO/1.4) was set to 100×/1.4 and the focal length of Fourier length was 300 mm. The optical parameters of microlens array were the same with 7-view settings.

The simulation parameters for the 3-view & 7-view system in Fig 2 were the same with above experimental settings, except for the 19-views one. 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.

Progressive sampling strategy

INR-based approach[22] adopted random sampling strategy to constant optimize the scene-representation network. The number of coordinates depended the speed of network convergence, which was multiplicatively increased with the size of 3D scene. Further, when applied such random sampling strategy to represent the 3D biological scene, the uneven distribution of grayscale intensity and structural complexity in fluorescent images made this representation difficult, denoted by the weak-signal loss

and slow convergence (Supplementary Figure 3). We attributed this deficiency to the ignoration of each voxel's contribution when conducting loss summation and averaging, implying that the network tended to repeatedly optimize the "easy" area with low-frequency and high intensity[43]. Inspired by the hard example mining in object detection[44], we developed a progressive sampling strategy to focus on the positions where contains structural distortion and signal loss.

The sampling strategy contains 3 steps: sparse sampling, full-sampling and focal sampling as shown in Fig 1b.

Firstly, for the optimization process of one scene, the captured FLFM views were down-sampled to half the size of original ones. Accordingly, the spatial meshgrid and the projection matrixes (FLFM PSF and geometry projection matrix) in hybrid physics model were also re-sampled to match the location and sampling rate of the resized views. Based on the low-frequency implicit bias[43] in neural network, this sparse sampling strategy simplified the fitting task on complex 3D signal and reduced the number of query voxels, thereby promoting the convergence at the initial stage.

Secondly, the neural network was optimized under the supervision of multi-views with the original sampling rate to recover the high-frequency details of the 3D scene, which would exist signal loss like random sampling strategy. Therefore, during the optimization process, an error map was simultaneously calculated and saved for the following step as the sampling probability map. Denoting $y' \subseteq \mathbb{R}^{N\times N}$ as the center-view projection from LFINR 3D prediction with geometry-model, and $y \subseteq \mathbb{R}^{N\times N}$ as the target enhanced center view. The error map can be formulated as:

$$ErrorMap(y', y) = \hat{1} - SSIMMap(y', y) - \frac{1}{N^2}\sum_{i=1}^{N^2} SSIMMap(y', y)$$

$$SSIMMap(y', y) = \frac{(2\mu_y\mu_{y'} + c_1)(2\sigma_{yy'} + c_2)}{(\mu_y^2 + \mu_{y'}^2 + c_1)(\sigma_y^2 + \sigma_{y'}^2 + c_2)}$$

where $\hat{1}$ means an all-ones matrix with the size of center view and N means the lateral size of the center view. μ_y and $\mu_{y'}$ are the mean of y and y', respectively. σ_y^2 and $\sigma_{y'}^2$ stand for the variance of y and y'. $c_1 = 0.01^2$ and $c_2 = 0.03^2$ are constants for 32-bit data (normalized to (0,1)). The kernel size used in SSIM calculation was set to 7. Then, we converted this map into a sampling probability map via dividing each pixel value by the summation of error map.

Thirdly, based on the calculated dynamic probability map, we conducted selective optimization on different regions of scene to recover the structural details. The time consumptions of various samples were listed in Supplementary Table 1.

Hybrid model rendering.

During the view rendering phase, LFINR adopted hybrid physics model to generate dual forward projection. The first term in the hybrid model was described by the FLFM PSF based on the wave-optics theory. In this paper, we modified the source code of oLaF[45] to generate PSF matrix, which was then convolved with represented 3D scene to produce wave-optics projection. The second term involved a geometric-optical model to describe the propagation process of ideal rays. Similar to the convolution

process in first part, this rendering process can be also described as the convolution between 3D volume and geometry projection matrix. To obtain this matrix, the 3D positions of maximum value of FLFM PSF at different depth were extracted first. Then, conducting linear fitting on these coordinates to obtain the depth-wise shift functions of different views:

$$x^i(z) = f_x^i(z)$$

$$y^i(z) = f_y^i(z)$$

where i is the view index and $x\&y\&z$ are the 3D spatial coordinates. f_x^i and f_y^i are the fitted parallax function, which revealed the projective geometry relation. Then, input all z-coordinates into these two functions to obtain the 3D positions of chief rays. Finally, an all-zero matrix with the size of the FLFM PSF was created, and voxels at the obtained 3D positions were assigned a value of 1.

Geometry consistency.

To assess the fidelity of 3D reconstructions and underscore the view-synthesis ability of LFINR, we employed geometry-consistency, computed from the structural similarity of Epi-polar Plane Images (EPIs)[27] generated from 3D reconstructions and ground truth, which can be formulated as:

$$Geometryconsistency = SSIM\big(EPI(y'), EPI(y)\big)$$

$$EPI(\cdot) = Reslice(Proj_{geometry}(\cdot))$$

where y' is the reconstruction result from deconvolution-based methods or LFINR and y is the ground truth. $Proj_{geometry}$ refers to the generation of views with horizontal arrangement by a tomography projection model. The number of generated views was set to 19 to ensure angular continuity. The angle range of this model was set to $(-\theta_{obj}, \theta_{obj})$ where θ_{obj} is the maximum acquisition angle decided by objective's NA. The operation of *Reslice* involves two steps: *a*. Concatenate the generated 2D projections along z-axis to obtain "view-stack"; *b*. "Top-reslice" the "view-stack" in Fiji to obtain axial section. This section plane is the EPI, containing both spatial information and angular information. The whole process is illustrated in Supplementary Figure 8.

Network implementation of LFINR

During one network optimization process, the input coordinates would follow 3 steps to obtain the computed loss: 1) Position encoding 2) MLP inference 3) Loss computation (Supplementary Figure 2).

Position encoding. The coordinates' low-frequency variation among 3D space would degrade the rendering quality of structural details or color variation[22]. Therefore, in LFINR, we adopt the Fourier positional encoding to convert such variation to the encoded high-frequency information:

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

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

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. The detailed settings of L_{xy} and L_z in various sample are listed in Supplementary Table 1. After position encoding, a high-dimensional feature vector would be obtained, denoting as $v \subseteq \mathbb{R}^{N\times L}$, where N is the number of input coordinates and L is the channel of encoded feature deiced by L_{xy} and L_z. Each element v_i in v can be expressed as:

$$v_i = [\gamma(p_{xy}), \gamma(p_z), x, y, z]$$

MLP inference. We utilized a Multi-layer Perceptron (MLP) to represent the 3D scene with querying spatial coordinates. MLP consists of input layer, hidden layers and output layers. The detailed parameters are seen in Supplementary Table 1. This process can be formulated as:

$$I_i = Relu(MLP(v_i))$$
$$Vol_{pred} = Reshape(\{I_i\}_{i=1}^{N})$$

where i is the index of some voxel and $I(i)$ is corresponding grayscale. $Relu$ is used to truncate the predicted values smaller than zero. After network inference, the collection of predicted voxels would be reshaped to the 3D-grid image (Vol_{pred}).

Loss computation. In LFINR, the loss function consisted two parts: L_2 loss function in wave-optics projection process and structural loss function in geometry projection process (Fig 1b). Specifically, the represented volume was convolved with FLFM PSF to fit the captured LF views and also with geometry projection matrix to fit enhanced views. The objective function of network optimization based on these views can be expressed as:

$$Loss = L_2(Vol_{pred} * PSF_o, y) + \lambda_1\left(1 - SSIM(Vol_{pred} * PSF_g, \hat{y})\right) + \lambda_2 \nabla_{zz}\left(Vol_{pred}\right) + \lambda_3 L_1(Vol_{pred})$$

where PSF_o and PSF_g are wave-optics PSF and geometry PSF, respectively. $\hat{y}$ and y are enhanced views and raw views. ∇_{zz} and $L_1(\cdot)$ are the regularization term to avoid model overfitting. λ_1, λ_2, λ_3 are weighting coefficients (Supplementary Table 1).

Data pre-processing

For static scene, the raw LF images underwent background subtraction by rolling-ball subtraction in Fiji. Subsequently, the mean intensity of each view was rescaled to be the same. Additionally, the PSF of each view was divided by the corresponding summation to conduct normalization. For dynamic scenes, in addition to background subtraction, the mean intensity of each frame was rescaled to be uniform. Due to the noise inherent in living cell imaging, a lightweight self-supervised denoising network (Supplementary Note 3) was constructed to enhance the signal-to-noise ratio (SNR) based on the captured LF videos. During the training stage, all input LF images were normalized to the range (0, 1).

Dynamic-scene representation by LFMNR

The essence of implicit neural representation (INR) is to represent complex high-dimensional signal smoothly with neuron network. For living biological imaging, the signal varies along 3D spatial coordinates and a 1D time axis, presenting challenges for network representation. One plausible approach is to add the 't-variable' directly into INR optimization to learn the 4D function $f(x, y, z, t)$. However, this approach significantly increases the computational burden due to the massive inputs of four-dimensional coordinates, especially for long-term imaging. Moreover, sudden grayscale changes caused by fast motion further complicate the four-dimensional function. Another approach involves employing an additional network to predict the motion modality of 3D signal[46], but constructing such motion model requires sophisticated handcrafting, making it challenging to accurately model irregular motion patterns and local morphologies of biological samples. Inspired by redundancy elimination techniques in data compression[47], we proposed a meta-learning-based approach to represent dynamic 3D scenes encoded in light field sequence, termed light-field meta neural representation (LFMNR), enabling the acceleration of network convergence based on high inter-frame correlation. Meta-learning, often referred to as "learning to learn", is a broad concept in metacognition that involves understanding and improving one's own learning processes. In this context, meta-learning refers to using optimization results from one scenario as the starting point for the next scenario's optimization, aiming to accelerate the learning process of neural networks. Specifically, during training phase, we first optimized network to represent the first volume with the guidance of input T_1 views, which produced network weights $\theta_1^{'}$ and optimizer Opt_1. The learning rate was 5×10^{-4} and the number of total iterations were 12000. For the second volume, the network and optimizer were initialized with parameters $\theta_1^{'}$ and with Opt_1, respectively. Then, the network was updated under the supervision of T_2 views. For remained volumes, according to consecutively transferring the previous parameters, the network convergence time was dramatically. During transfer-learning stage, the number of iterations was 100 for each time-point, which cost ~20 s.

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. Guo, C., Liu, W., Hua, X., Li, H. & Jia, S. Fourier light-field microscopy. *Opt Express* **27**, 25573-25594 (2019).
9. Hua, X., Liu, W. & Jia, S. High-resolution Fourier light-field microscopy for volumetric multi-color live-cell imaging. *Optica* **8**, 614-620 (2021).
10. Hua, X. et al. Light-field flow cytometry for high-resolution, volumetric and multiparametric 3D single-cell analysis. *Nat. Commun.* **15**, 1975 (2024).
11. Frikel, J. & Quinto, E.T. Characterization and reduction of artifacts in limited angle tomography. *Inverse Problems* **29**, 125007 (2013).
12. Kak, A.C. & Slaney, M. Principles of computerized tomographic imaging. (SIAM, 2001).
13. Wang, Z. et al. Real-time volumetric reconstruction of biological dynamics with light-field microscopy and deep learning. *Nat. Methods* **18**, 551-556 (2021).
14. Zhu, L. 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).
15. Wagner, N. et al. Deep learning-enhanced light-field imaging with continuous validation. *Nat. Methods* **18**, 557-563 (2021).
16. 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).
17. Yi, C. et al. Video-rate 3D imaging of living cells using Fourier view-channel-depth light field microscopy. *Communications Biology* **6**, 1259 (2023).
18. Belthangady, C. & Royer, L.A. Applications, promises, and pitfalls of deep learning for fluorescence image reconstruction. *Nat. Methods* **16**, 1215-1225 (2019).
19. Wang, F. et al. Far-field super-resolution ghost imaging with a deep neural network constraint. *Light: Science & Applications* **11**, 1 (2022).
20. Wang, F. et al. Phase imaging with an untrained neural network. *Light: Science & Applications* **9**, 77 (2020).
21. Huang, L., Chen, H., Liu, T. & Ozcan, A. Self-supervised learning of hologram

reconstruction using physics consistency. *Nature Machine Intelligence* **5**, 895-907 (2023).
22. Mildenhall, B. et al. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM* **65**, 99-106 (2021).
23. Yuan, Y.-J. et al. in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 18353-18364 (2022).
24. Li, Z. et al. in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 8456-8465 (2023).
25. 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).
26. Chen, A. et al. in Proceedings of the IEEE/CVF International Conference on Computer Vision 14124-14133 (2021).
27. Stankiewicz, O., Lafruit, G. & Domański, M. in Academic Press Library in Signal Processing, Volume 6. (eds. R. Chellappa & S. Theodoridis) 3-74 (Academic Press, 2018).
28. Mo, Y. et al. Quantitative structured illumination microscopy via a physical model-based background filtering algorithm reveals actin dynamics. *Nat. Commun.* **14**, 3089 (2023).
29. Culley, S. et al. Quantitative mapping and minimization of super-resolution optical imaging artifacts. *Nat. Methods* **15**, 263-266 (2018).
30. Frank, S. et al. The role of dynamin-related protein 1, a mediator of mitochondrial fission, in apoptosis. *Developmental cell* **1**, 515-525 (2001).
31. Wasilewski, M. & Scorrano, L. The changing shape of mitochondrial apoptosis. *Trends in Endocrinology & Metabolism* **20**, 287-294 (2009).
32. 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).
33. Chen, B. et al. Projective light-sheet microscopy with flexible parameter selection. *Nat. Commun.* **15**, 2755 (2024).
34. Pearl, N., Treibitz, T. & Korman, S. in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 12672-12681 (2022).
35. 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).
36. Rantala, M. et al. Wavelet-based reconstruction for limited-angle X-ray tomography. *IEEE transactions on medical imaging* **25**, 210-217 (2006).
37. Lyu, Q., Neph, R. & Sheng, K. Tomographic detection of photon pairs produced from high-energy X-rays for the monitoring of radiotherapy dosing. *Nature Biomedical Engineering* **7**, 323-334 (2023).
38. Turk, M. & Baumeister, W. The promise and the challenges of cryo-electron tomography. *FEBS letters* **594**, 3243-3261 (2020).
39. Zheng, G., Horstmeyer, R. & Yang, C. Wide-field, high-resolution Fourier ptychographic microscopy. *Nature photonics* **7**, 739-745 (2013).
40. Zhou, H. et al. FPM-INR: Fourier ptychographic microscopy image stack reconstruction using implicit neural representations. *arXiv preprint arXiv:2310.18529* (2023).
41. Fessler, J.A. Model-Based Image Reconstruction for MRI. *IEEE Signal Processing Magazine* **27**, 81-89 (2010).

42. Shen, L., Pauly, J. & Xing, L. NeRP: implicit neural representation learning with prior embedding for sparsely sampled image reconstruction. *IEEE Transactions on Neural Networks and Learning Systems* (2022).
43. Xu, Z.-Q.J., Zhang, Y., Luo, T., Xiao, Y. & Ma, Z. Frequency principle: Fourier analysis sheds light on deep neural networks. *arXiv preprint arXiv:1901.06523* (2019).
44. Shrivastava, A., Gupta, A. & Girshick, R. in Proceedings of the IEEE conference on computer vision and pattern recognition 761-769 (2016).
45. Stefanoiu, A., Scrofani, G., Saavedra, G., Martínez-Corral, M. & Lasser, T. What about computational super-resolution in fluorescence Fourier light field microscopy? *Opt Express* **28**, 16554-16568 (2020).
46. 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).
47. Strümpler, Y., Postels, J., Yang, R., Gool, L.V. & Tombari, F. in European Conference on Computer Vision 74-91 (Springer, 2022).