Research Article Vol. 34, No. 5 / 9 Mar 2026 / Optics Express 7429
Optics EXPRESS

Seeing through fibers: unsupervised image reconstruction in fiber bundle imaging systems

AMIR REZA VAZIFEH,[1] CONGLI WANG,[2] AMOGH JOSHI,[1] ILYA CHUGUNOV,[2] JIPENG SUN,[2] JIWOON YEOM,[2] JASON W. FLEISCHER,[1] JOSÉ S. PULIDO,[3] AND FELIX HEIDE[2,*]

[1]*Department of Electrical and Computer Engineering, Princeton University, Princeton, New Jersey 08544, USA*
[2]*Department of Computer Science, Princeton University, Princeton, New Jersey 08544, USA*
[3]*Wills Eye Hospital, Philadelphia, Pennsylvania 19107, USA*
[*]*fheide@princeton.edu*

Abstract: Fiber bundle imaging systems suffer from sampling artifacts such as honeycomb patterns due to their discrete and non-uniform fiber layout, fundamentally limiting image resolution. Conventional reconstruction methods rely on precise calibration of the fiber layout or learning from paired datasets, both of which have limited generalization across imaging setups and require sample-specific preparation. We present an unsupervised method for reconstructing high-resolution images using a burst of misaligned frames that does not require known fiber layout, paired training data, or per-sample calibration. Our approach jointly solves motion estimation and image reconstruction through test-time training. We model each burst frame as a deformed observation of a single canonical view, parameterizing the underlying motion with a coordinate-based network. A second coordinate-based network learns a joint super-resolved scene representation shared across aligned frames. Both networks are trained jointly end-to-end without paired ground truth or external supervision. Simulation and experimental results demonstrate that our method robustly removes fiber bundle artifacts and generalizes to various sample types. We also released a benchmark dataset for optical fiber bundle imaging to facilitate future research.

1. Introduction

Fiber bundle imaging systems are widely used in endoscopy [1–6], optical coherence tomography [7,8], fluorescence imaging [9,10], and neuronal and cellular imaging [11,12]. Each fiber in the bundle transmits light from a small region, resulting in a pixelated image where the number and arrangement of fibers limit the lateral resolution. This introduces characteristic sampling artifacts, such as honeycomb patterns and intensity discontinuities between fiber cores. Furthermore, the cladding between the fibers blocks incoming light, leaving gaps in spatial information, degrading image fidelity [13]. Such artifacts obscure fine structures, making image interpretation difficult and motivating computational methods for high-resolution image reconstruction.

Several image reconstruction methods have been proposed using a single image as input. Supervised learning is a common approach, in which models are trained using experimental data (ground truth from dual-sensor systems) [14], or synthetic data generated by simulating fiber artifacts [15,16], to map distorted images to clean counterparts. However, reliance on ground truth makes supervised learning impractical for clinical or *in vivo* settings, and models trained in specific domains fail to generalize to new imaging conditions, sample types, or fiber configurations. Other methods have leveraged prior information about the distortion sources to guide reconstruction. For example, by estimating the fiber layout via imaging a white reference and using the known core positions, high-resolution images can be reconstructed using interpolation [17–19], maximizing an image prior [20], compressed sensing [21], or the iterative shrinkage thresholding algorithm [22]. While these methods do not require pairs of distorted and

#587675 https://doi.org/10.1364/OE.587675
Journal © 2026 Received 30 Dec 2025; revised 28 Jan 2026; accepted 6 Feb 2026; published 23 Feb 2026

Optics EXPRESS

clean images, they rely on precise manual calibration of fiber geometry, which can drift over time. Moreover, rather than truly recovering spatial details, they in-paint unobserved regions based on prior assumptions, limiting their ability to capture fine details. Recent work has explored untrained neural networks for fiber bundle imaging [23]; where fibers deliver time-varying speckle patterns rather than directly transmit spatial images, and does not target the spatial artifacts (honeycomb patterns, inter-core gaps) of fiber bundle imaging.

Capturing multiple frames can reveal additional information about regions that are obscured in individual images but accessible across multiple observations. However, this approach introduces the challenge of determining how content across frames should be aligned and merged. Several studies have addressed this by using clues about the fiber layout, such as the distance between fiber cores, to design controlled motion (e.g., shifts or rotations of the imaging probe or fiber bundle) to capture image bursts [18,24–27]. To merge frames into a high-resolution image, techniques like interpolation [18,24,25,28] and image processing followed by up-sampling [26] or averaging [27] are used. Instead of relying on controlled motion, inter-frame transformations can be estimated from the data. For example, Shao et al. [29] modeled each low-resolution capture as a warped, fiber core–filtered, noisy observation of the high-resolution scene. They formulated an inverse problem to reconstruct the high-resolution image by estimating geometric transformations and maximizing a posterior with a smoothing prior. In a follow-up work, they developed a supervised deep learning framework that first aligns raw fiber bundle sequences using a motion estimation network, then applies a 3D convolutional neural network (CNN) to map aligned sequences to ground truth images [30]. Despite promising results, most prior work relies on accessing ground truth, fiber layout, or motion prior, and is validated on limited samples.

Implicit neural representations (INRs) are an emerging approach for modeling signal values as functions of spatial or temporal coordinates parameterized by multilayer perceptrons (MLPs) [31]. Unlike traditional discrete representations (e.g., pixel grids), INRs provide continuous signal representations at arbitrary resolution. INRs have shown success in various applications, including image and video synthesis [32], novel 3D view generation [33–35], tomography [36], depth estimation [37], and hyperspectral imaging [38,39]. INRs have been extended to unsupervised layer separation tasks, such as de-moiréing or fence removal using image bursts [40,41]. Inspired by these recent advances, we introduce a fully unsupervised framework for removing fiber bundle artifacts using misaligned image bursts acquired under arbitrary motion. Our method jointly learns to align frames and reconstruct a clean canonical view by optimizing two coordinate-based MLPs end-to-end. The first MLP estimates deformation fields to align the input frames by learning parameters of a homography transformation, while the second synthesizes a high-resolution image by modeling the shared scene content across aligned views. Unlike previous methods that require ground truths [14–16,30], access to the fiber layout [17–22], or prior assumptions about motion [24–27], our data-driven approach is fully unsupervised, calibration-free, and broadly applicable across imaging conditions and sample types. We evaluate our method on both simulated and experimental datasets and demonstrate significant spatial resolution enhancement.

2. Methods

INRs provide a continuous functional representation of discrete signals using MLPs. For images, they represent a function mapping spatial coordinates to RGB values ($\mathbb{R}^2 \rightarrow \mathbb{R}^3$), and for videos, a function mapping spatio-temporal coordinates ($\mathbb{R}^3 \rightarrow \mathbb{R}^3$), enabling continuous interpolation across space and time. In fiber-bundle imaging, individual frames suffer from occluded fiber cores that obscure different spatial regions. However, due to relative motion between frames, occluded information in one frame may be visible in others. This motivates using INRs to jointly model motion and scene content: by learning transformations that align frames, we can aggregate complementary information across the temporal sequence to reconstruct the complete scene.

Optics EXPRESS

2.1. Reconstruction algorithm

Figure 1 summarizes our imaging setup and reconstruction method. Moving objects are imaged through a lens, relayed by a fiber bundle, and magnified by secondary optics. Our method removes fiber bundle artifacts from a burst of misaligned frames using two components: a motion estimation network that represents inter-frame transformations, and a scene representation network. The reconstructed RGB value $\hat{I}$ at coordinate (x, y) in frame t is given by:

$$\hat{I}(x, y, t) = [\hat{R}, \hat{G}, \hat{B}] = f_\theta\big(\gamma(T_{g_\phi}(x, y, t))\big). \tag{1}$$

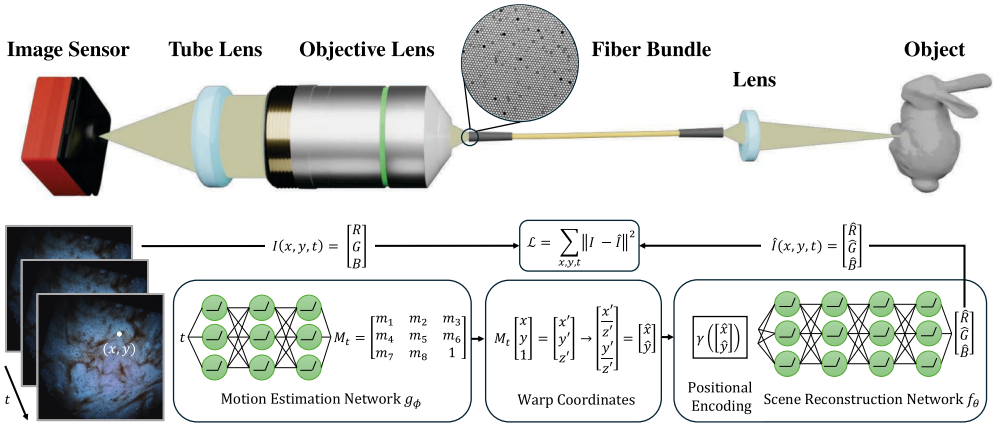

Fig. 1. Overview of the imaging setup and reconstruction algorithm. The goal is to remove fiber-bundle artifacts by observing the scene across multiple frames and merging information without using ground truth. The first network, g_ϕ, aligns the frames by predicting homography parameters, while the second network, f_θ, reconstructs the scene from the aligned frames.

Here, T_{g_ϕ} represents a spatial transformation (e.g., homography or optical flow) parameterized by network g_ϕ. The warped coordinates $(\hat{x}, \hat{y}) = T_{g_\phi}(x, y, t)$ are fed into the scene representation network f_θ, which maps them to RGB values. Both networks are MLPs, which struggle to learn high-frequency components of a scene, a limitation known as spectral bias [42,43]. To mitigate this, warped coordinates $(\hat{x}, \hat{y}) = T_{g_\phi}(x, y, t)$ are first passed through a positional encoding $\gamma(\cdot)$. Networks g_ϕ and f_θ are trained jointly using the Adam optimizer [44] (learning rate 10^{-4} with default momentum parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$ for 3000 iterations) to minimize the reconstruction loss given by

$$\mathcal{L} = \sum_{x,y,t} \left\| \hat{I}(x, y, t) - I(x, y, t) \right\|^2, \tag{2}$$

where $I(x, y, t)$ denotes the observed RGB value at coordinate (x, y) in frame t. After training, the artifact-free scene at frame t is reconstructed by applying Eq. (1) at desired spatio-temporal coordinates (x, y, t).

2.2. Ablation study

Since our method has two stages – (1) the MLP g_ϕ for motion estimation and frame alignment, and (2) the MLP f_θ for scene reconstruction – we ablate each stage and our design choices to evaluate their impact on reconstruction quality. To determine the impact of the motion model, we fix the scene reconstruction network f_θ and positional encoding $\gamma(.)$, then test different motion models,

Optics EXPRESS

including a no-motion baseline, optical flow (with/without total variation (TV) regularizer), and homography. To evaluate the effect of the reconstruction network f_θ and positional encoding $\gamma(.)$, we fix the motion model to homography and vary only f_θ's architecture and encoding. Specifically, we tested ReLU-MLP without positional encoding, with NeRF positional encoding [33], with Fourier positional encoding [32], with Hash positional encoding [45], and sinusoid activation function without positional encoding [46]. These positional encoding strategies aim to enhance the network's ability to capture fine details [32].

2.2.1. Impact of motion model

We compare three motion models: no-motion, optical flow, and homography. The scene reconstruction network f_θ and positional encoding $\gamma(.)$ are set to ReLU with Fourier positional encoding [32] in all experiments.

No Motion: For no motion compensation, the goal is to reconstruct the scene by memorizing RGB values at each spatiotemporal coordinate using f_θ. With a fixed fiber mask, the model may recover occluded regions if they appear in other frames. The model and loss are defined as:

$$\hat{I}(x,y,t) = [\hat{R},\hat{G},\hat{B}] = f_\theta(\gamma(x,y,t)), \quad \mathcal{L}_{\text{no-motion}} = \sum_{x,y,t} \left\| \hat{I}(x,y,t) - I(x,y,t) \right\|^2. \tag{3}$$

Optical Flow: Optical flow aims to estimate the location of a pixel in frame $t+1$ given its position (x,y) in frame t. Motion is represented as a pixel-wise displacement $(\Delta x_t, \Delta y_t)$, implicitly assuming $I(x,y,t) \approx I(x+\Delta x_t, y+\Delta y_t, t+1)$. The motion is predicted by a neural network $g_\phi(x,y,t)$, which is used to warp the input coordinates before feeding them to the reconstruction network f_θ. Both networks are trained jointly, with an additional TV loss to encourage smooth motion fields.

$$\begin{aligned} (\Delta x_t, \Delta y_t) = g_\phi(x,y,t), \quad (\hat{x},\hat{y}) &= (x+\Delta x_t, y+\Delta y_t), \\ \hat{I}(x,y,t) = [\hat{R},\hat{G},\hat{B}] &= f_\theta(\gamma(\hat{x},\hat{y})), \\ \mathcal{L}_{\text{optical-flow}} = \sum_{x,y,t} \left\| \hat{I}(x,y,t) - I(x,y,t) \right\|^2 &+ \lambda \cdot \text{TV}(g_\phi). \end{aligned} \tag{4}$$

Homography: Instead of estimating motion for each pixel, homography-based alignment tries to find a global transformation for each frame. Specifically, the transformation at frame t is represented by a homography matrix M_t, predicted by a network $g_\phi(t)$. This matrix warps the coordinates (x,y) in frame t, and the network f_θ uses these warped coordinates to reconstruct the scene:

$$\begin{aligned} M_t = g_\phi(t) = \begin{bmatrix} m_1 & m_2 & m_3 \\ m_4 & m_5 & m_6 \\ m_7 & m_8 & 1 \end{bmatrix}, \quad \begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = M_t \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}, \quad (\hat{x},\hat{y}) = \left(\frac{x'}{z'}, \frac{y'}{z'} \right), \\ \hat{I}(x,y,t) = [\hat{R},\hat{G},\hat{B}] = f_\theta(\gamma(\hat{x},\hat{y})), \quad \mathcal{L}_{\text{homography}} = \sum_{x,y,t} \left\| \hat{I}(x,y,t) - I(x,y,t) \right\|^2. \end{aligned} \tag{5}$$

Each motion model offers different trade-offs between reconstruction quality and robustness. According to Eq. (4), optical flow estimates motion at the pixel level, so the motion network g_ϕ must predict a displacement vector for every pixel in every frame. For example, for a 20-frame video at a pixel resolution of 1200×1200, g_ϕ must estimate approximately 28.8 million spatio-temporal correspondences across frames. This makes optimization more challenging, requires a deeper network for training, and increases computational costs. Optical flow is also inherently sensitive to self-similar and textureless structures as it relies on the brightness constancy and local appearance matching across frames. When similar patterns repeat spatially

Optics EXPRESS

or regions lack distinctive texture, local neighborhoods lead to unreliable motion estimation. To achieve stable and smooth flows, additional regularization such as TV is often required [47]. In contrast, homography assumes global planar motion and applies a single transformation across the entire frame. Specifically, for T frames at resolution $H \times W$, g_ϕ for homography is a mapping from $\mathbb{R} \rightarrow \mathbb{R}^8$ (frame index t to homography parameters in Eq. (5)) that should be found only for T frames, while for optical flow is a mapping from $\mathbb{R}^3 \rightarrow \mathbb{R}^2$ that should be found for all $H \times W \times T$ spatio-temporal coordinates. This makes homography more robust and easier to optimize, though it cannot model scenes with large depth variations between frames or parallax.

Figure 2 compares the different motion models on two experimental samples (additional examples in Supplement 1). The degradation of reconstruction quality in the no-motion baseline highlights the importance of motion modeling. Optical flow without TV regularization fails to reconstruct the scene and produces visible artifacts. Even with TV regularizer, optical flow struggles with the textureless or self-similar structures: in the first example, fiber patterns remain visible in the green textureless regions, and in the second example, the reconstruction fails due to the highly repetitive tower structure. In contrast, homography achieves significantly better visual quality and faster convergence in both scenarios.

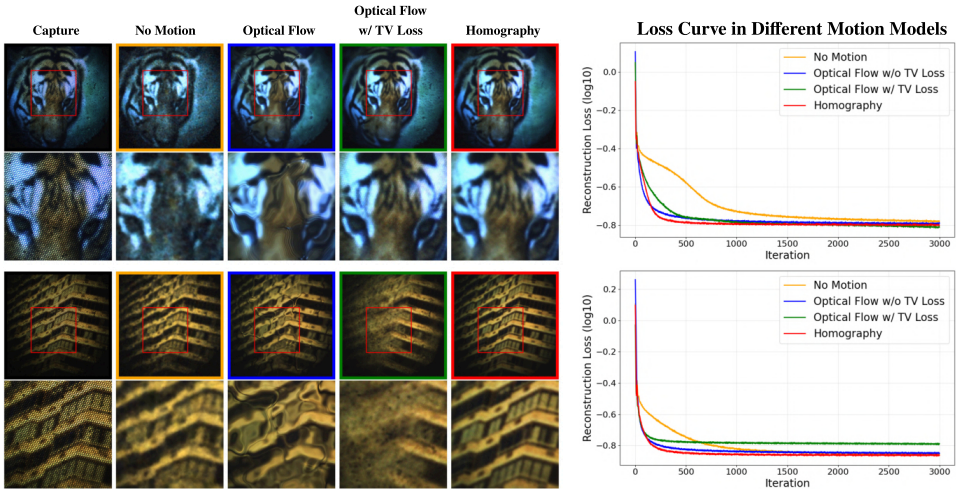

Fig. 2. Comparison of motion models on reconstruction quality. The scene reconstruction MLP f_θ is set to ReLU with Fourier positional encoding $\gamma(.)$ in all cases, while different motion models g_ϕ are tested. For each video, the first frame is shown.

2.2.2. Impact of MLP architecture and positional encoding

Assuming homography as the motion model, the warped coordinates $(\hat{x}, \hat{y})$ at frame t are computed by Eq. (5). The warped coordinates $(\hat{x}, \hat{y})$ can be fed into the reconstruction MLP f_θ with different positional encodings. We now ablate the effects of different MLP architectures and positional encodings $\gamma(.)$ as follows:

ReLU MLP without Positional Encoding: The most direct approach is to input the raw warped coordinates to a standard MLP with ReLU activations, $\hat{I}(x, y, t) = f_\theta([\hat{x}, \hat{y}]; t)$.

ReLU MLP with NeRF Positional Encoding [33]: The coordinates can be mapped to a higher-dimensional Fourier feature space using a positional encoding function $\gamma(\hat{x}, \hat{y})$, inspired

Optics EXPRESS

by NeRF, with logarithmically spaced frequencies; The mapping becomes:

$$\gamma(\hat{x},\hat{y}) = \Big[\hat{x}, \sin(2^0\pi\hat{x}), \cos(2^0\pi\hat{x}), \ldots, \sin(2^L\pi\hat{x}), \cos(2^L\pi\hat{x}), \\ \hat{y}, \sin(2^0\pi\hat{y}), \cos(2^0\pi\hat{y}), \ldots, \sin(2^L\pi\hat{y}), \cos(2^L\pi\hat{y})\Big], \tag{6}$$

where L is a hyperparameter controlling the number of frequency bands. It is tuned according to the complexity of the scene; smaller L for smooth images, larger L for fine details. The reconstruction network predicts $\hat{I}(x,y,t) = f_\theta(\gamma(\hat{x},\hat{y});t)$.

ReLU MLP with Fourier Positional Encoding [32]: Instead of logarithmically spaced frequencies, a set of random Fourier features is sampled from a Gaussian distribution. The warped coordinates are projected through a random Gaussian matrix $\mathbf{B} \in \mathbb{R}^{m\times 2}$ with entries $b_{ij} \sim \mathcal{N}(0,\sigma^2)$, yielding the positional encoding:

$$\gamma(\hat{x},\hat{y}) = \Big[\hat{x}, \sin\big(2\pi\,\mathbf{b}_1^\top \begin{bmatrix}\hat{x}\\ \hat{y}\end{bmatrix}\big), \ldots, \sin\big(2\pi\,\mathbf{b}_m^\top \begin{bmatrix}\hat{x}\\ \hat{y}\end{bmatrix}\big), \\ \hat{y}, \cos\big(2\pi\,\mathbf{b}_1^\top \begin{bmatrix}\hat{x}\\ \hat{y}\end{bmatrix}\big), \ldots, \cos\big(2\pi\,\mathbf{b}_m^\top \begin{bmatrix}\hat{x}\\ \hat{y}\end{bmatrix}\big)\Big], \tag{7}$$

where $\mathbf{b}_i \in \mathbb{R}^2$ is the i-th row of $\mathbf{B}$. In practice, m and σ are tuned according to the complexity of the scene: smaller values capture smooth variations, larger values allow finer details. The exact distribution family is less critical than its standard deviation, which controls the frequency range that the network can learn. The network predicts $\hat{I}(x,y,t) = f_\theta(\gamma(\hat{x},\hat{y});t)$.

ReLU MLP with Hash Positional Encoding [45]: Unlike NeRF [33] or Fourier encoding [32] that rely on fixed sinusoidal functions, multi-resolution hash encoding maps coordinates to learnable high-dimensional vectors via a cascade of hash tables across multiple resolutions. This approach is adaptive, automatically prioritizing fine-scale details, and is highly optimized. Its main hyperparameters are the number of levels, and the number of features per level, which control the resolution hierarchy and feature dimensionality. The reconstruction network then uses $\hat{I}(x,y,t) = f_\theta(\gamma(\hat{x},\hat{y});t)$.

SIREN MLP without Positional Encoding [46]: Instead of using a separate positional encoding with a ReLU MLP, an MLP with sinusoidal activation functions (SIREN) can be applied directly to raw coordinates to capture fine details. The SIREN MLP uses a frequency factor ω_0 in the first layer, treated as a hyperparameter and tuned according to scene complexity. The reconstruction is given by $\hat{I}(x,y,t) = f_\theta^{\text{siren}}([\hat{x},\hat{y}];t)$.

Each architecture presents different trade-offs in reconstruction quality and convergence speed. The reconstruction for two examples using different positional encodings is shown in Fig. 3 (more examples in Supplement 1). Our experiments show that Fourier encoding achieves the best reconstruction quality, followed by SIREN. NeRF positional encoding struggles to capture high-frequency details and converges more slowly than the first two. ReLU MLP without positional encoding fails to reconstruct the scene, highlighting the importance of positional encoding. Hash encoding improves over the original captures, but still exhibits the honeycomb effect of the fiber. Remarkably, Fourier features and SIREN have the fastest convergence speed, followed by hash encoding, NeRF, and finally ReLU without positional encoding.

Optics EXPRESS

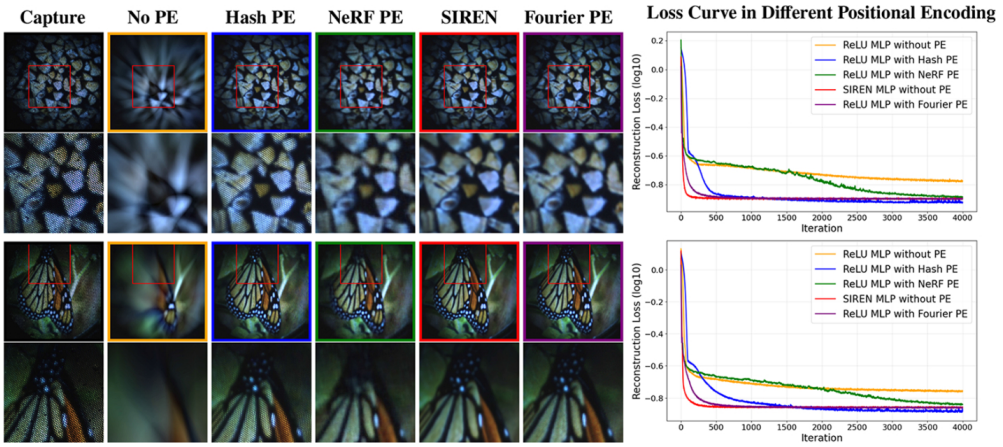

Fig. 3. Comparison of MLP architectures and positional encodings for the scene reconstruction MLP f_θ. The motion estimation MLP g_ϕ is fixed to a homography model for all cases. For each video, the first frame is shown.

3. Imaging setup

Figure 4 shows our fiber bundle imaging setup. The scene is imaged by a lens with a field of view of (FOV) approximately 38° onto one end of a SCHOTT fiber bundle (NA 0.38, acceptance angle 45°, 7.6 µm core diameter, approximately 18,000 fibers). The wide-angle endoscope lens (about 1.9 mm diameter, 120° FOV) is a commercial off-the-shelf component for coupling into fiber bundles with similar numerical apertures, ensuring efficient light collection across the image field. The other end of the fiber bundle is relayed to a 1/1.2" color sensor (FLIR, BFLY-U3-23S6C-C) with 1920 × 1200 pixels and a pixel pitch of 5.86 µm via relay optics consisting of an objective (Nikon, CFI Plan Fluor, NA 0.30, working distance 16 mm with 10× magnification, FOV 2.5 mm) and an achromatic doublet (Thorlabs, AC508-150-A-ML, focal length 150 mm) as a tube lens for chromatic aberration correction across the visible spectrum.

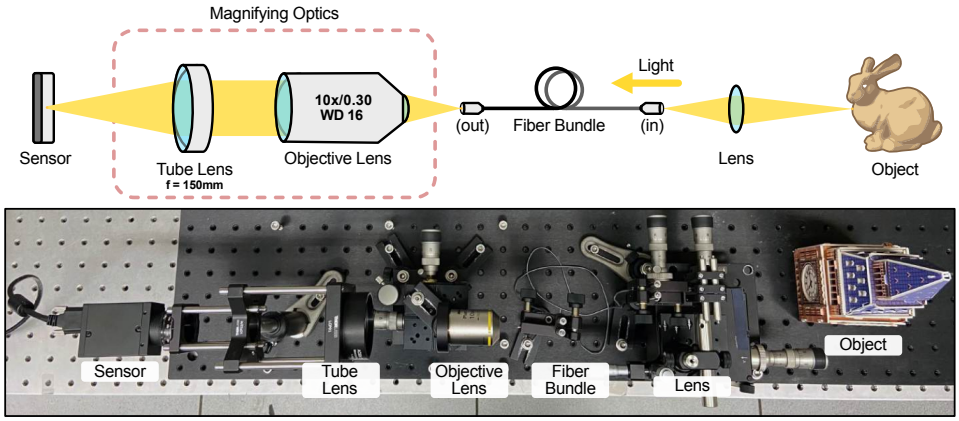

Fig. 4. Fiber bundle imaging system employing a lens (38° FOV) coupled to a SCHOTT fiber bundle that relays the object image through a 10× Nikon objective lens and achromatic doublet to a color image sensor. The relay optics are aligned to maximize the highest image contrast at the fiber bundle output, with the 3D object positioned within the depth of field of the lens. For 2D experimental data acquisition, the object was substituted with a screen displaying test images.

Optics EXPRESS

4. Results

By modeling motion with homography and using a ReLU MLP with Fourier positional encoding as the scene representation network (with the frequency scale $\sigma = 5$ used as the default value), we show results on both simulated and experimental data. We then analyze the impact of the number of frames, evaluate robustness to random motion, and report the algorithm's runtime.

4.1. Evaluation in simulation

To validate our approach, we simulated fiber-bundle burst images on five samples: histological images of the liver and kidney, lens tissue, the resolution chart, and a tiger image. We then applied random Brownian motion to generate 20-frame videos. The results, showing the first frame of each video, are presented in Fig. 5 (details and additional examples in Supplement 1). We compared our reconstruction method against Gaussian filtering, a supervised CNN trained on paired simulated data and tested on unseen samples, and PyFiberBundle [18]. We also compared our method to INR-based inpainting, where an MLP is trained on known core intensities to reconstruct the unknown ones, as described in Supplement 1. Our reconstruction removes the honeycomb pattern and recovers fine structures obscured in the simulated images: individual cells and nuclei in liver and kidney tissue, fine fibrous structures in lens tissue, and fine line patterns in the resolution chart. On average, our method achieved a PSNR improvement of 11.1 dB and an SSIM improvement of 0.41 across the five samples compared to the simulated capture. Detailed statistics for each sample and comparative results against all baseline methods are reported in Table 1.

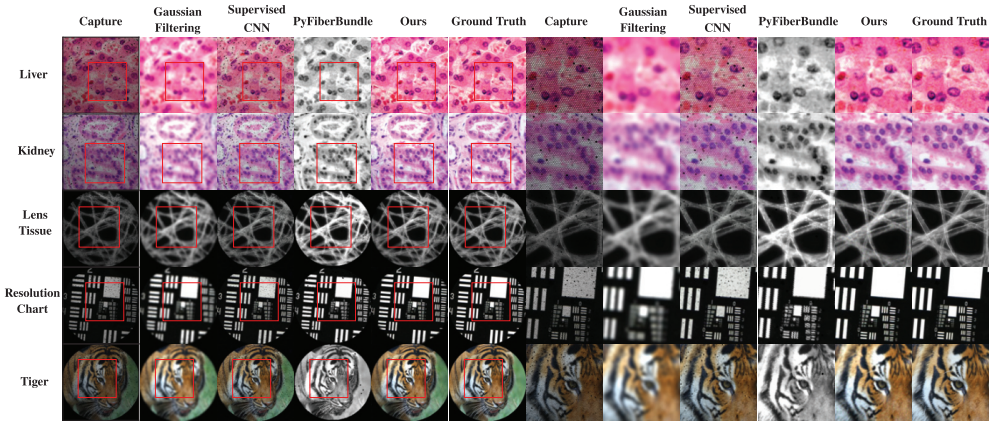

Fig. 5. Reconstruction results on simulated data. First frames from simulated captures and corresponding reconstructions are displayed for each scene. Captured videos are available in Dataset 1 [48].

Table 1. Quantitative reconstruction results for the samples in Fig. 5.

Video Name	Video Dimension	PSNR [dB]/SSIM ↑						Runtime (s) ↓	Compression ↑
		Captured	Gaussian Filtering	Supervised CNN	PyFiberBundle [18]	Ours	Improvement		
Liver	(20,600,600,3)	8.66/0.15	20.23/0.53	14.57/0.25	18.84/**0.70**	**22.63**/0.67	+13.97/+0.52	48.05	90.81%
Kidney	(20,580,580,3)	7.92/0.10	21.34/0.68	14.51/0.27	21.41/0.75	**22.26/0.75**	+14.34/+0.65	35.26	90.81%
Lens Tissue	(20,944,944,3)	14.89/0.46	23.25/0.73	21.34/0.76	13.99/0.64	**23.54/0.81**	+8.65/+0.35	93.80	96.91%
Resolution Chart	(20,1000,1000,3)	13.02/0.63	15.78/0.66	19.52/0.79	13.80/0.72	**23.13/0.85**	+10.11/+0.22	102.93	96.91%
Tiger	(20,1000,1000,3)	13.17/0.39	21.08/0.67	19.91/0.71	15.28/0.66	**21.83/0.72**	+8.66/+0.33	103.51	96.91%

Optics EXPRESS

4.2. Experimental evaluation (2D scene)

For experimental validation, we displayed a set of 2D images on a screen and applied random motion to generate a 20-frame video. These frames were then imaged using our setup. The reconstruction results for six examples are shown in Fig. 6 (more examples in Supplement 1). Our reconstruction reveals structural details that are not visible in the experimental captures, such as white spots on the orange butterfly wings and vascular structures in the tissue sample. Specifically, the reconstruction of the resolution chart reaches Group 0 Element 4 versus Group 0 Element 1 in the raw capture, yielding a ≈1.41× resolution improvement.

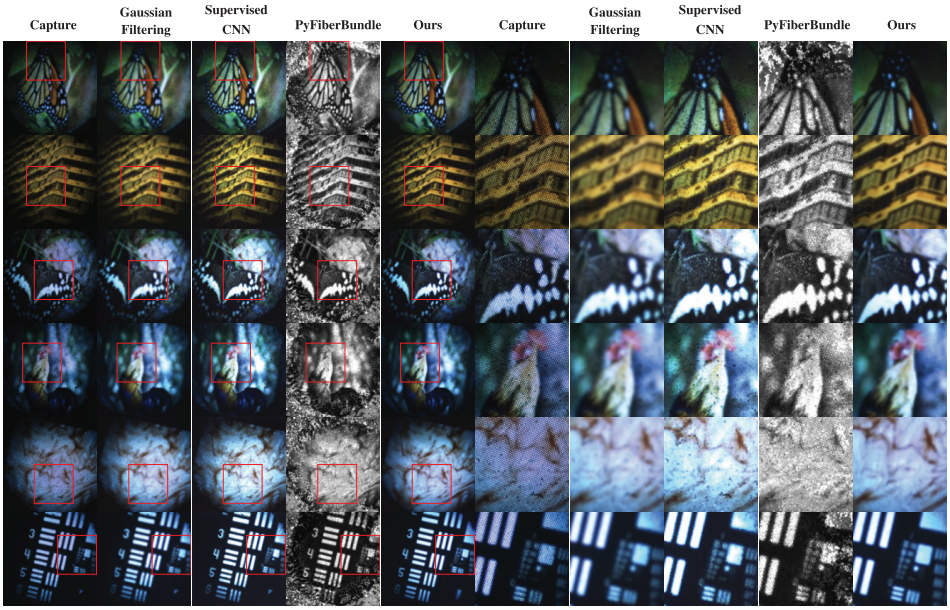

Fig. 6. Reconstruction results on 2D scenes displayed on a screen and captured with our imaging setup. For each video, the first frame is shown (Available in Dataset 1 [48]).

To evaluate robustness under different fiber configurations, we applied our method to an image burst of the 1951 USAF resolution chart from [18], and we report resolution in pixels. The 5-pixel fiber core pitch imposes a Nyquist limit of 10 pixels/line-pair; the raw image already exceeds this, resolving Group 7 Element 1 (4.3 pixels/lp). As shown in Figure 7, our reconstruction further extends resolution to Group 7 Element 5 (2.7 pixels/lp), a ≈1.59× improvement over raw. This result is achieved without any prior knowledge of the imaging setup or fiber configuration, and relies solely on a burst of 12 misaligned frames. In contrast, PyFiberBundle [18], which requires prior knowledge of the fiber layout and is limited to grayscale images, failed on our data.

4.3. Experimental evaluation (3D object)

To further evaluate performance in real-world conditions, we imaged three-dimensional objects using our imaging setup. Random motion was applied by manually moving objects in front of the imaging setup, producing videos of 20 frames for reconstruction. The reconstruction results are shown in Fig. 8. As before, our method resolves details invisible in the individual captures, such as the fine tower structures, the spokes of the truck wheel, and the surface texture of the duck.

However, in the presence of parallax or when the scene contains objects at different depths, the homography-based alignment may cause the entire learned scene to move together in the reconstructed video, resulting in apparent shearing or translation of the image rather than correct

Optics EXPRESS

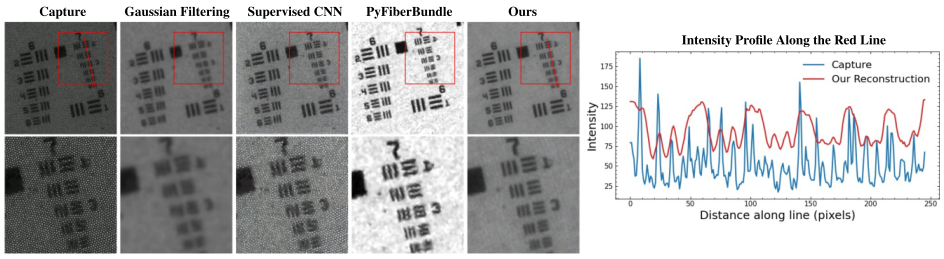

Fig. 7. Reconstruction results on a burst of 12 frames from [18]. For each video, only the first frame is shown. The intensity profile along the red line is shown on the right.

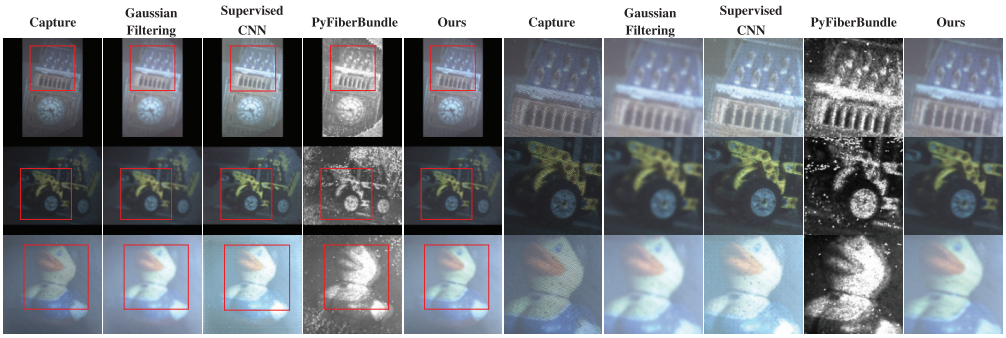

Fig. 8. Reconstruction results on 3D objects captured with our imaging setup. Random motion was applied by moving the objects manually during capture. For each video, the first frame is shown (Available in Dataset 1 [48]).

alignment of individual objects. If the depth variation or parallax is significant, more advanced motion models or additional metadata may be required for proper scene modeling.

4.4. Influence of number of frames

Since our method relies on processing an image burst, we expect the reconstruction quality to improve as more frames are used, at the cost of longer test-time optimization and potentially deeper MLPs. This is because missed spatial information can be observed in subsequent frames, enabling more accurate reconstructions. Figure 9 shows the images reconstructed using subsets of the first $\{1, 2, 5, 10, 20\}$ frames of a 20-frame burst. Increasing the number of frames leads to progressively finer details and reduced mosaic artifacts. This improvement arises from multi-frame densification and gap filling, rather than surpassing the optical limits of the system. Importantly, the same MLP architecture, motion model, and number of iterations are used in all cases, implying that our framework can accommodate varying numbers of frames.

4.5. Robustness to random frame order and motion

Unlike previous studies that rely on prior knowledge of motion or carefully designed motion patterns based on the distances between the fiber cores [24–27], our approach can reconstruct scenes directly from an image burst under arbitrary motion. We validated this through two examples: (1) experimental data, where frame orders were randomly shuffled (last column of the upper two rows in Fig. 9), and (2) simulation data, where random Brownian motion was generated and used for scene reconstruction (last row in Fig. 9).

Optics EXPRESS

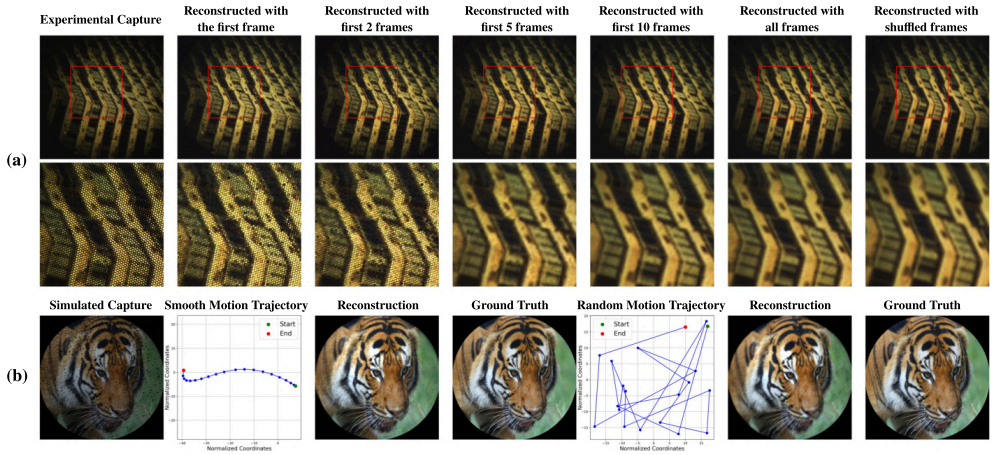

Fig. 9. Reconstruction quality as a function of the number of input frames and motion randomness. (a) Experimental results: increasing the number of frames improves image detail and resolution. The last column demonstrates robustness when the 20 input frames are reordered, showing the algorithm's ability to handle random motion. (b) Simulation results: two different motion trajectories generate image bursts, illustrating how motion randomness affects reconstruction. For each video, the first frame is shown.

4.6. *Compressed representation of video*

The networks g_ϕ and f_θ can be used not only to reconstruct super-resolved video but also as an alternative representation of it, as video coordinates (x, y, t) are used as input values to the MLPs. For example, an RGB video with 20 frames at a resolution of 1200×1200 contains 86,400,000 uint8 values (1 byte per value, i.e. ~82.4 MB), while the two MLPs provide an equivalent representation with only 464,139 float32 learnable parameters (4 bytes per parameter, i.e. ~1.77 MB), achieving a compression ratio ~97.85% (More details of the MLP in Supplement 1). In other words, our method achieves super-resolved reconstruction while simultaneously offering a highly compressed representation of the video. The compression ratio for five simulation samples is provided in Table 1, ranging from 90.80% to 96.91% depending on the depth and width of the MLP, as well as the dimensions of the video.

4.7. *Reconstruction runtime*

For experimental data, processing a 20-frame RGB video at 1200×1200 resolution requires approximately 2 minutes on an NVIDIA A100 GPU for 3000 iterations. More results for simulation data are provided in Table 1, with run times ranging from 48 to 104 seconds.

Optics EXPRESS

Table 2. Comparison of image reconstruction methods in fiber bundle imaging systems.

Work	Year	Input	Ground Truth Required?	Fiber Layout Required?	Motion Prior?	Code & Data Availability	Method
[29]	2018	Burst	No	Yes	No	No	Forward model includes fiber PSF and core localization. Images aligned via homography, then inverse problem solved using MAP estimation with Laplacian smoothness prior.
[17]	2018	Image	No	Yes	N/A	No	Localization of fiber cores by imaging a white reference with Delaunay triangulation and interpolation to fill gaps
[14]	2019	Image	Yes	No	N/A	No	Supervised Generative Adversarial Network (GAN) Trained on paired noisy–clean experimental data
[30]	2019	Burst	Yes	No	No	No	Train separate supervised networks: one estimates each frame's homography for alignment, the other merges aligned frames using a 3D CNN.
[27]	2020	Burst	No	No	Yes	No	Rotate bundle in a controlled way to capture multiple frames and merge them via image processing
[18]	2023	Both	No	Yes	No	Yes	Image processing functions using triangular interpolation and spatial filtering, mostly requiring fiber core spacing and/or fiber layout pattern
[15]	2025	Image	Yes	Yes	N/A	No	GAN trained on paired noisy–clean synthetic data and tested on experimental captures
Ours	2026	Burst	No	No	No	Yes	Joint optimization of motion estimation and scene reconstruction using coordinate-based networks

Optics EXPRESS

5. Discussion

Self/unsupervised approaches have been used for image reconstruction in different imaging modalities such as single-photon imaging [49] and multimode fibers (MMFs) [13,50]; however, their application to fiber bundle imaging remains unexplored. Here, we propose an end-to-end unsupervised framework for image reconstruction in fiber bundle imaging using multiple misaligned frames. Our method jointly trains two networks: one predicts homography parameters for each frame, and the other reconstructs the underlying scene using a coordinate-based network. The closest prior work [30] also used two separate networks: one for alignment and a 3D CNN for mapping aligned frames to a high-resolution image. However, both networks are trained separately in a supervised manner (requiring paired ground truth). The total number of frames must be known in advance and cannot differ from the number used during training. Models trained on grayscale cannot handle RGB or hyperspectral images, and alignment assumes the first frame as the reference. In contrast, our method does not require ground truth data, motion priors, or knowledge of fiber layout, making it more flexible and generalizable across different acquisition conditions. It supports arbitrary numbers of input frames and naturally extends to grayscale, RGB, and hyperspectral imaging with no modification, as the scene reconstruction network predicts channel-wise intensities at continuous spatio-temporal coordinates. Unlike [30], our homography estimation does not assume any frame as a reference. To better position our contribution relative to prior work, we provide a comprehensive comparison in Table 2.

Despite these advantages, several limitations remain. Reconstruction speed is currently the primary bottleneck, with 20-frame bursts requiring up to two minutes to process. Significant acceleration could be achieved through custom CUDA operators for MLP training [45]. Although the method effectively recovers the information masked by the fiber, even better resolution could be achieved by deconvolving the point spread function of the imaging system [51]. Optimal reconstruction depends on empirically tuned positional encoding parameters, but it might not be sufficient for scenes with spatially varying frequency content. For these cases, we can run algorithm on smaller crop of scene where the frequency of underlying scene is more consistent. Additionally, the homography-based motion model may fail in scenarios with significant depth variation or parallax between frames, motivating the adoption of more expressive motion models or the use of metadata from the imaging setup. Hardware limitations, arising from the low light efficiency of the fiber optics and the performance of the camera sensor, could be alleviated by using more light-sensitive components [52].

Future developments should target expanded FOV and hyperspectral imaging capabilities beyond RGB channels. In particular, our framework can increase the effective FOV by capturing a burst of misaligned frames, aligning them with homography, and synthesizing a single wide-FOV reconstruction. It can also be extended to hyperspectral imaging by replacing the RGB output with multi-wavelength intensity profiles. Learning implicit representations of scenes with spatially varying frequencies without scene-specific positional encoding tuning is an active area of research [53,54], and its connection to physical parameters like fiber core spacing is an interesting future direction. Evaluation of our algorithm on clinical datasets, including *in vivo* imaging, is left for future work. Although our method focuses on fiber bundle imaging systems, multimode fibers represent a complementary direction. Unlike fiber bundle imaging, which consists of thousands of individual fiber cores, MMFs employ a single core through which all light propagates, enabling smaller probe diameters and are therefore attractive for minimally invasive imaging, but they produce speckle-like outputs that require fundamentally different reconstruction strategies. Recent unsupervised learning approaches for MMF image reconstruction have been proposed [13,50], but they require sample-type-specific calibration and remain limited in general applicability. Extending coordinate-based neural networks to MMF imaging is left for future work.

Optics EXPRESS

6. Conclusion

In summary, we propose an unsupervised test-time learning method to reconstruct super-resolved images in fiber bundle imaging systems from an image sequence. The key idea is to recover information occluded by the fiber mask in individual frames but revealed across multiple frames. Our method jointly learns to align frames and reconstruct a clean canonical view by optimizing two coordinate-based MLPs end-to-end. The first MLP estimates the motion to align the input frames, for which we adopt a homography-based model. The second MLP synthesizes a high-resolution image by modeling shared scene content across multiple frames, with its positional encoding controlling the level of fine detail that can be reconstructed. Each of these design choices, the motion model and the positional encoding, can be adapted to scene content and may also have their own specific hyperparameters. Experimental and simulation results show that our method can significantly improve spatial resolution without requiring any ground truth, prior knowledge of the fiber layout, or clues about the underlying motion. We also release a publicly available dataset to provide a benchmark for future research.

Funding. NSF CAREER Award (2047359); Packard Foundation Fellowship; Sloan Research Fellowship; Sony Young Faculty Award; Project X Innovation Award; Amazon Science Research Award; Bosch Research Award; Air Force Office of Scientific Research (FA9550-23-1-0221, FA9550-21-1-0317).

Acknowledgment. The authors thank Ethan Tseng for valuable discussions. Felix Heide is a co-founder of Algolux (now Torc Robotics), Head of AI at Torc Robotics, and a co-founder of Cephia AI.

Disclosures. The authors declare that there are no conflicts of interest related to this article.

Data availability. Code and dataset underlying the results presented in this paper, including all experimental captures from the fiber bundle imaging setup and simulated datasets, are publicly available in Ref. [48].

Supplemental document. See Supplement 1 for supporting content.

References

1. Z. Liu, L. Wang, Y. Meng, *et al.*, "All-fiber high-speed image detection enabled by deep learning," Nat. Commun. **13**(1), 1433 (2022).
2. A. Perperidis, K. Dhaliwal, S. McLaughlin, *et al.*, "Image computing for fibre-bundle endomicroscopy: A review," Med. Image Anal. **62**, 101620 (2020).
3. D. Kim, J. Moon, M. Kim, *et al.*, "Toward a miniature endomicroscope: pixelation-free and diffraction-limited imaging through a fiber bundle," Opt. Lett. **39**(7), 1921–1924 (2014).
4. K. Vyas, M. Hughes, and G.-Z. Yang, "Electromagnetic tracking of handheld high-resolution endomicroscopy probes to assist with real-time video mosaicking," in *Endoscopic Microscopy X; and Optical Techniques in Pulmonary Medicine II*, vol. 9304M. J. Suter, S. Lam, M. Brenner, *et al.*, eds. (SPIE, 2015), p. 93040Y.
5. J. Sun, B. Zhao, D. Wang, *et al.*, "Calibration-free quantitative phase imaging in multi-core fiber endoscopes using end-to-end deep learning," Opt. Lett. **49**(2), 342–345 (2024).
6. M. Hughes, T. P. Chang, and G.-Z. Yang, "Fiber bundle endocytoscopy," Biomed. Opt. Express **4**(12), 2781–2794 (2013).
7. J.-H. Han, J. Lee, and J. U. Kang, "Pixelation effect removal from fiber bundle probe based optical coherence tomography imaging," Opt. Express **18**(7), 7427–7439 (2010).
8. J.-H. Han, S. M. Yoon, and G.-J. Yoon, "Decoupling structural artifacts in fiber optic imaging by applying compressive sensing," Optik **126**(19), 2013–2017 (2015).
9. A. Porat, E. R. Andresen, H. Rigneault, *et al.*, "Widefield lensless imaging through a fiber bundle via speckle correlations," Opt. Express **24**(15), 16835–16855 (2016).
10. B. A. Flusberg, E. D. Cocker, W. Piyawattanametha, *et al.*, "Fiber-optic fluorescence imaging," Nat. Methods **2**(12), 941–950 (2005).
11. M. Pierce, D. Yu, and R. Richards-Kortum, "High-resolution fiber-optic microendoscopy for *in situ* cellular imaging," Journal of Visualized Experiments: JoVE p. e2306 (2011).
12. N. Farah, A. Levinsky, I. Brosh, *et al.*, "Holographic fiber bundle system for patterned optogenetic activation of large-scale neuronal networks," Neurophotonics **2**(4), 045002 (2015).
13. X. Hu, J. Zhao, J. E. Antonio-Lopez, *et al.*, "Unsupervised full-color cellular image reconstruction through disordered optical fiber," Light: Sci. Appl. **12**(1), 125 (2023).
14. J. Shao, J. Zhang, X. Huang, *et al.*, "Fiber bundle image restoration using deep learning," Opt. Lett. **44**(5), 1080–1083 (2019).
15. J. Chen, W. Shang, and S. Xu, "Endoir: A GAN-based method for fiber bundle endoscope image restoration," Opt. Lasers Eng. **184**, 108588 (2025).

Optics EXPRESS

16. E. Kim, S. Kim, M. Choi, *et al.*, "Honeycomb artifact removal using convolutional neural network for fiber bundle imaging," Sensors **23**(1), 333 (2022).
17. P. Wang, G. Turcatel, C. Arnesano, *et al.*, "Fiber pattern removal and image reconstruction method for snapshot mosaic hyperspectral endoscopic images," Biomed. Opt. Express **9**(2), 780–790 (2018).
18. M. R. Hughes, "Real-timing processing of fiber bundle endomicroscopy images in Python using PyFibreBundle," Appl. Opt. **62**(34), 9041–9050 (2023).
19. K. Vyas, M. Hughes, B. G. Rosa, *et al.*, "Fiber bundle shifting endomicroscopy for high-resolution imaging," Biomed. Opt. Express **9**(10), 4649–4664 (2018).
20. J.-H. Han and S. M. Yoon, "Depixelation of coherent fiber bundle endoscopy based on learning patterns of image prior," Opt. Lett. **36**(16), 3212–3214 (2011).
21. S. P. Mekhail, N. Abudukeyoumu, J. Ward, *et al.*, "Fiber-bundle-basis sparse reconstruction for high resolution wide-field microendoscopy," Biomed. Opt. Express **9**(4), 1843–1851 (2018).
22. X. Liu, L. Zhang, M. Kirby, *et al.*, "Iterative ℓ_1-min algorithm for fixed pattern noise removal in fiber-bundle-based endoscopic imaging," J. Opt. Soc. Am. A **33**(4), 630–636 (2016).
23. Y. Zhang, J. Xue, Z. Jiang, *et al.*, "Efficient untrained neural network for high resolution optical fiber bundle compressive imaging," Opt. Laser Technol. **192**, 114079 (2025).
24. K. Vyas, M. Hughes, and G.-Z. Yang, "Fiber-shifting endomicroscopy for enhanced resolution imaging," in *Frontiers in Optics 2017*, (OSA, 2017), p. JTu2A.79.
25. C.-Y. Lee and J.-H. Han, "Elimination of honeycomb patterns in fiber bundle imaging by a superimposition method," Opt. Lett. **38**(12), 2023–2025 (2013).
26. G. W. Cheon, J. Cha, and J. U. Kang, "Random transverse motion-induced spatial compounding for fiber bundle imaging," Opt. Lett. **39**(15), 4368–4371 (2014).
27. C. Renteria, J. Suárez, A. Licudine, *et al.*, "Depixelation and enhancement of fiber bundle images by bundle rotation," Appl. Opt. **59**(2), 536–544 (2020).
28. Y. Huang, W. Zhou, B. Xu, *et al.*, "Resolution improvement in real-time and video mosaicing for fiber bundle imaging," OSA Continuum **4**(10), 2577 (2021).
29. J. Shao, W.-C. Liao, R. Liang, *et al.*, "Resolution enhancement for fiber bundle imaging using maximum a posteriori estimation," Opt. Lett. **43**(8), 1906–1909 (2018).
30. J. Shao, J. Zhang, R. Liang, *et al.*, "Fiber bundle imaging resolution enhancement using deep learning," Opt. Express **27**(11), 15880–15890 (2019).
31. K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," Neural Networks: The Off. J. Int. Neural Netw. Soc. **2**(5), 359–366 (1989).
32. M. Tancik, P. Srinivasan, B. Mildenhall, *et al.*, "Fourier features let networks learn high frequency functions in low dimensional domains," in *Advances in Neural Information Processing Systems*, (2020), pp. 7537–7547.
33. B. Mildenhall, P. P. Srinivasan, M. Tancik, *et al.*, "NeRF: representing scenes as neural radiance fields for view synthesis," Commun. ACM **65**(1), 99–106 (2022).
34. V. Sitzmann, M. Zollhöfer, and G. Wetzstein, "Scene Representation Networks: Continuous 3D-structure-aware neural scene representations," in *Advances in Neural Information Processing Systems*, (2019).
35. A. Pumarola, E. Corona, G. Pons-Moll, *et al.*, "D-NeRF: Neural radiance fields for dynamic scenes," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (IEEE, 2021), pp. 10318–10327.
36. Y. Sun, J. Liu, M. Xie, *et al.*, "CoIL: Coordinate-based internal learning for tomographic imaging," IEEE Trans. Comput. Imaging **7**, 1400–1412 (2021).
37. I. Chugunov, Y. Zhang, and F. Heide, "Shakes on a plane: Unsupervised depth estimation from unstabilized photography," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (IEEE, 2023), pp. 13240–13251.
38. J. Boondicharern, A. R. Vazifeh, and J. W. Fleischer, "Snapshot hyperspectral imaging via compressive sensing and implicit neural representation," in *Optica Imaging Congress 2025 (3D, DH, COSI, IS, pcAOP,RadIT)*, (Optica Publishing Group, 2025), p. CTu1B.5.
39. R. Ma and S. He, "Multi-channel volume density neural radiance field for hyperspectral imaging," Sci. Rep. **15**(1), 16253 (2025).
40. I. Chugunov, D. Shustin, R. Yan, *et al.*, "Neural spline fields for burst image fusion and layer separation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (IEEE, 2024), pp. 25763–25773.
41. S. Nam, M. A. Brubaker, and M. S. Brown, "Neural image representations for multi-image fusion and layer separation," in *European Conference on Computer Vision, vol. abs/2108.01199 of Lecture Notes in Computer Science* (Springer Nature Switzerland, 2022), pp. 216–232.
42. N. Rahaman, A. Baratin, D. Arpit, *et al.*, "On the spectral bias of neural networks," in *International Conference on Machine Learning*, (2019).
43. R. Basri, M. Galun, A. Geifman, *et al.*, "Frequency bias in neural networks for input of non-uniform density," in *International Conference on Machine Learning*, (2020), pp. 64:1–64:10.
44. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv (2014).
45. T. Müller, A. Evans, C. Schied, *et al.*, "Instant neural graphics primitives with a multiresolution hash encoding," ACM Trans. Graph. **41**(4), 1–15 (2022).
46. V. Sitzmann, J. N. P. Martel, A. W. Bergman, *et al.*, "Implicit neural representations with periodic activation functions," in *Advances in Neural Information Processing Systems*, (2020).

Optics EXPRESS

47. S. Baker, D. Scharstein, J. P. Lewis, *et al.*, "A database and evaluation methodology for optical flow," Int. J. Comput. Vis. **92**(1), 1–31 (2011).
48. A. Reza Vazifeh, "Neural field fiber bundle imaging," GitHub (2026), [Accessed Feb 11] https://github.com/princeton-computational-imaging/Neural-Field-Fiber-Bundle-Imaging.
49. Y. Chen, C. Jiang, and Y. Pan, "Single-photon image super-resolution via self-supervised learning," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, (IEEE, 2023), pp. 1–5.
50. C. Zhang, Y. Shi, Z. Yao, *et al.*, "Imaging through multimode fibers using physics-assisted unsupervised learning," Opt. Lasers Eng. **194**, 109183 (2025).
51. F. Heide, M. Rouf, M. B. Hullin, *et al.*, "High-quality computational imaging through simple lenses," ACM Trans. Graph. **32**(5), 1–14 (2013).
52. J. E. Fröch, L. Huang, Q. A. A. Tanguy, *et al.*, "Real time full-color imaging in a Meta-optical fiber endoscope," eLight **3**(1), 13 (2023).
53. V. Saragadam, D. LeJeune, J. Tan, *et al.*, "WIRE: Wavelet implicit neural representations," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (IEEE, 2023), pp. 18507–18516.
54. Z. Liu, H. Zhu, Q. Zhang, *et al.*, "FINER: Flexible spectral-bias tuning in implicit NEural representation by variableperiodic activation functions," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (IEEE, 2024), pp. 2713–2722.