Recover Biological Structure from Sparse-View Diffraction Images with Neural Volumetric Prior

Supplementary Material

Overview

This supplementary material consists of 7 sections that provide extended data and insights into the Neural Volumetric Prior (NVP). First, detailed equations describing the rendering process based on diffraction optics are provided. The second section elaborates on the coherent alignment for processing experimentally captured images. The third section discusses the optical section of FDT using k-space analysis. The fourth section specifies the loss functions. The fifth section explains evaluation metrics used to design and optimize NVP. The sixth section provides implementation details, including the optimization of feature dimensions and network layers. The seventh section presents an ablation study on self-calibration.

Furthermore, four supplementary tables are included to present extended results. Table 4 compares the reconstruction performance of various methods across different numbers of measurements, extending Table 2. Table 5 examines the effect of feature dimensionality on NVP's performance. Table 6 evaluates the impact of network depth on NVP's performance. Table 7 provides an ablation study investigating the role of self-calibration for experimentally captured data.

1. Multi-slice Model for Rendering Process

To accurately model light propagation through a 3D heterogeneous semi-transparent sample, the sample is represented by multiple slices overlapping in the light propagation direction, so called the "multi-slice model". To improve computational efficiency, we extend the conventional multi-slice model from classical optimization algorithms (i.e., FISTA) [17] to the PyTorch framework. This differentiable model allows automated backpropagation. It also incorporates a coarse-to-fine strategy for adjusting parameters and fine-tuning the RI.

In this model, the 3D phase object is represented as a stack of N_z layers, each with an unknown RI, $\hat{n}_k(\mathbf{r})$ for $k = 1, 2, ..., N_z$, where:

$$\hat{n} \triangleq \{\hat{n}_k(\mathbf{r})\}_{k=1}^{N_z}, \quad \mathbf{r} = (x, y), \tag{4}$$

As light propagates through each layer, its phase is altered according to the transmission function $t_k(\mathbf{r})$:

$$t_k(\mathbf{r}) = \exp\left(\frac{j2\pi}{\lambda}\Delta z(\hat{n}_k(\mathbf{r}) - n_b)\right), \tag{5}$$

where λ is the wavelength, Δz is the layer thickness, and n_b is the background RI. The Fresnel propagation operator $\mathcal{P}_{\Delta z}$ represents light propagation through the layers:

$$\mathcal{P}_{\Delta z}\{\cdot\} = \mathcal{F}^{-1}\left\{\exp\left(-j2\pi\Delta z\sqrt{\left(\frac{1}{\lambda}\right)^2 - ||\mathbf{k}||^2}\right)\cdot\mathcal{F}\{\cdot\}\right\}, \tag{6}$$

where $\mathcal{F}\{\cdot\}$ and $\mathcal{F}^{-1}\{\cdot\}$ are the Fourier transform and its inverse. The electric field $\hat{E}_{k,i}(\mathbf{r})$ propagating through the k-th layer for the i-th fluorescent source is:

$$\hat{E}_{k,i}(\mathbf{r}) = \mathcal{P}_{\Delta z}\left\{t_k(\mathbf{r})\cdot\hat{E}_{k-1,i}(\mathbf{r})\right\}. \tag{7}$$

At the image plane, the electric field is calculated by applying the pupil function $p(\mathbf{k})$:

$$\hat{E}_i(\mathbf{r}) = \mathcal{F}^{-1}\left\{p(\mathbf{k})\cdot\mathcal{F}\left\{\hat{E}_{N_z,i}(\mathbf{r})\right\}\right\}. \tag{8}$$

For thick objects, the field is back-propagated over a distance ΔZ_c before reaching the camera:

$$\hat{E}_i(\mathbf{r}) = \mathcal{F}^{-1}\left\{p(\mathbf{k})\cdot\mathcal{F}\left\{\mathcal{P}_{-\Delta Z_c}\left\{\hat{E}_{N_z,i}(\mathbf{r})\right\}\right\}\right\}. \tag{9}$$

The camera detects the intensity of the light field as:

$$\hat{I}_i(\mathbf{r}) = |\hat{E}_i(\mathbf{r})|^2. \tag{10}$$

2. Coherent Alignment

The multi-slice model allows us to generate synthetic images using rendering equations under the assumption that the illumination source is coherent. Coherence implies that all photons emitted from the source share the same frequency and phase. However, in the experimental data captured by FDT, the light source (i.e., fluorescence) is partially coherent, introducing model mismatching.

One approach to address this mismatch would be to mathematically model [81] the partially coherent sources to generate more accurate synthetic images. However, this approach is impractical due to the inherent randomness and spatially-variant distribution of fluorescence light.

To overcome this limitation, we propose a coherent alignment method (Figure 8). The synthetic image $\hat{I}$ (top left), generated using coherent light sources, is adjusted to match partial coherence by applying a coherent mask derived from the diffraction pattern (top center). This results

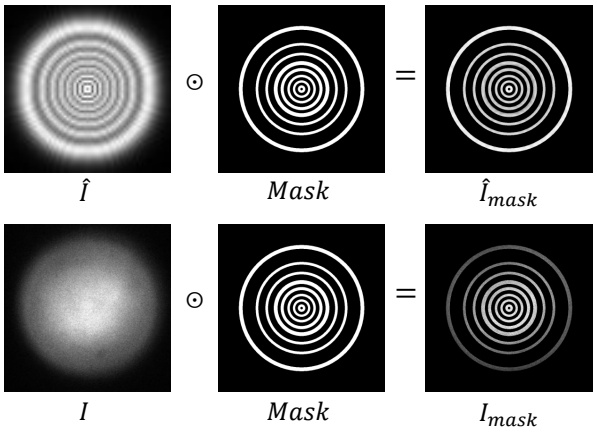

Figure 8. **Illustration of the coherent alignment method to address partial coherence in fluorescence-based illumination.** The top row shows the synthetic image $\hat{I}$ (left) multiplied by the coherent mask (center) to produce the masked synthetic image $\hat{I}_{\text{mask}}$ (right). The bottom row depicts the real image I (left) undergoing the same masking process to generate the masked real image I_{mask} (right). This approach aligns the partially coherent fluorescence illumination with the coherent multi-slice model, mitigating mismatches caused by variations in photon frequency and phase.

in the masked synthetic image $\hat{I}_{\text{mask}}$ (top right). A similar process is applied to the real observed image I (bottom left), which undergoes masking with the same coherent mask (bottom center) to produce the masked real image I_{mask} (bottom right). The loss is calculated between I_{mask} and $\hat{I}_{\text{mask}}$ instead of I and $\hat{I}$, which reduces model mismatches.

3. Missing cone problem.

NVP solves the missing cone problem by using fluorescence as a partially coherent light source for illumination. It is well established that partially coherent light can solve this issue [82]. We also provide the optical transfer function (OTF) of partially coherent illumination (Figure 9), which indicates that it fills the missing cone compared to traditional coherent illumination, thereby enabling optical sectioning for 3D reconstruction.

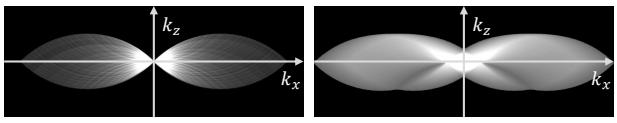

Figure 9. **OTF under coherent (left) and partially coherent (right) conditions.**

4. Loss

The image reconstruction loss $\mathcal{L}_{\text{img}}$ is a weighted sum of L_1, L_2, and Structural Similarity Index (SSIM) losses, ensuring robustness to pixel-wise differences while preserving perceptual quality. This loss is defined as:

$$\mathcal{L}_{\text{img}} = \frac{1}{N}\sum_{i=1}^{N} \alpha\|\hat{I}_i - I_i\|_1 + \beta\|\hat{I}_i - I_i\|_2^2 + \gamma\,\ell_{\text{SSIM}}(\hat{I}_i, I_i), \tag{11}$$

where N is the number of images, $\hat{I}_i$ and I_i are the predicted and ground-truth images, and α, β, and γ are the corresponding weights for each loss term.

Since the RI distribution in biological samples varies gradually, total variation regularization is applied across both the X-Y plane and the Z-axis to enforce smoothness in the predicted 3D RI. This is formulated as:

$$\mathcal{R}_{\text{ri}} = \lambda_{xy}\,TV_{xy}(\hat{n}) + \lambda_z\,TV_z(\hat{n}), \tag{12}$$

where $TV_{xy}(\hat{n})$ and $TV_z(\hat{n})$ represent the total variation for predicted RI $\hat{n}$ along the X-Y plane and Z-axis, respectively. The weights λ_{xy} and λ_z control the strength of regularization in each direction, the $\alpha, \beta, \gamma = 4, 4, 1.5$..

5. Metrics

We use three metrics to evaluate the quality of our 3D reconstructions: SSIM, Learned Perceptual Image Patch Similarity (LPIPS), and Peak Signal-to-Noise Ratio (PSNR). These metrics provide a comprehensive evaluation: SSIM for structural similarity, LPIPS for perceptual quality, and PSNR for signal fidelity. For evaluation, we normalize te image intensities to a range of $0-1$, while preserving the original reconstructed RI values, as they retain their real-world physical significance.

5.1. SSIM

SSIM measures the similarity between two images in terms of luminance, contrast, and structure, with values ranging from 0 to 1 (higher is better). The SSIM between two images x and y is defined as:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{13}$$

where μ_x and μ_y are the means, σ_x^2 and σ_y^2 are the variances, and σ_{xy} is the covariance of x and y. C_1 and C_2 are constants to stabilize the division.

5.2. LPIPS

LPIPS quantifies perceptual similarity by comparing high-level feature representations from a pre-trained neural network. Lower LPIPS values indicate greater similarity to the ground truth:

$$\text{LPIPS}(x, y) = \sum_l w_l\|\phi_l(x) - \phi_l(y)\|_2 \tag{14}$$

where ϕ_l represents features extracted from layer l of the neural network, and w_l is a weight for layer l.

5.3. PSNR

PSNR measures the ratio between the maximum possible signal power and the power of noise, with higher values indicating better quality:

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{L^2}{\text{MSE}} \right) \tag{15}$$

where L is the maximum pixel value (1 for normalized images) and MSE is the mean squared error between the reconstructed and ground truth images.

6. Implementation Details

In the main text, we compare explicit, triplane, and NVP methods. All methods are tested under the same experimental settings, including identical input images, random seeds, learning rates, and regularization methods.

For the explicit method, we directly adopt the parameter configurations from [18], as it operates on predefined grids without requiring neural network parameters. For the triplane and NVP methods, both the feature dimension and the neural network size must be experimentally determined. Tables 5 and 6 summarize the metrics used to evaluate the performance of NVP across various configurations with the synthetic cell dataset.

Feature Dimension Selection: Table 5 summarizes the performance across different feature dimensions. Increasing the feature dimension generally improves the SSIM and PSNR up to a dimension of 16, where the highest SSIM (0.9317) are achieved. Beyond 16, both SSIM and PSNR exhibit a decline. LPIPS, on the other hand, increases gradually as feature dimensions grow, highlighting a trade-off between feature complexity and perceptual quality. Considering these trends, the feature dimension of 16 offers the optimal balance, delivering superior SSIM and competitive performance in other metrics. This makes it the most effective choice for robust reconstruction.

Network Layers Selection: Table 6 evaluates the performance of NVP with different numbers of neural network layers. Similar to feature dimensions, adding layers initially improves performance, with the best results of PSNR achieved at 6 layers.

In summary, by experimentally determining these parameters, we ensure that NVP delivers state-of-the-art performance with optimal computational efficiency, making it a practical choice for 3D RI reconstruction of biological samples.

7. Ablation Study on Self-Calibration

Accurate calibration of viewpoints is critical in 3D rendering using the multi-slice model, as the precision of 3D reconstruction depends heavily on the exact geometry of each captured image. However, in experimental setups, light scattering and system inaccuracies often result in imprecise measurements of angles and positions. We develop a self-supervised calibration method to accurately determine illumination positions. In our experiments, the camera position is fixed while the illumination position is variable, whereas in natural scenarios, the camera position (viewpoint) typically changes while the illumination is fixed. Thus, in our method, calibrating the illumination positions is analogous to calibrating viewpoints. To evaluate the robustness of the self-calibration, we conducted an ablation study by introducing Gaussian noise to the illumination positions, representing the viewpoints.

In the ablation study, we simulated a real-world scenario by adding Gaussian noise to the illumination positions. The noise had a mean of 0, a standard deviation of 0.01, and a maximum value of 0.05, relative to the illumination location range of $[-0.5, 0.5]$. This represents a significant perturbation, as the noise amplitude is substantial compared to the total range of viewpoint values. We then compared the reconstruction quality of the synthetic tissue dataset with and without self-calibration under these noisy conditions.

Table 7 shows the results of this ablation study. Self-calibration demonstrates substantial improvements in reconstruction metrics for both RI and IMG. For RI, self-calibration reduces MSE from 1.28×10^{-2} to 5.71×10^{-3}, increases SSIM from 0.2989 to 0.3911, improves LPIPS from 0.8297 to 0.6732, and raises PSNR from 18.9189 to 22.4331. For IMG, self-calibration maintains high SSIM (0.9117 compared to 0.9871 without calibration) and achieves a lower LPIPS (0.0540 compared to 0.0133), while keeping PSNR stable.

In summary, these results emphasize the importance of self-supervised calibration in mitigating large misalignment in illumination positions and achieving robust and accurate 3D reconstruction, even under challenging conditions with substantial perturbations to the viewpoint locations.

Number	Method	Data	MSE↓	SSIM↑	LPIPS↓	PSNR↑
5	nvp	RI	$\mathbf{6.9800 \times 10^{-2}}$	**0.4238**	**0.5140**	**11.5615**
		IMG	3.0475×10^{-7}	0.9999	0.0000	65.1605
	exp	RI	7.9224×10^{-2}	0.3623	0.5777	11.0114
		IMG	9.4795×10^{-5}	0.9857	0.0176	40.2322
	tri	RI	1.3439×10^{-1}	0.0484	0.7143	8.7162
		IMG	1.5973×10^{-5}	0.9969	0.0029	47.9662
7	nvp	RI	$\mathbf{6.8529 \times 10^{-2}}$	**0.4775**	**0.5038**	**11.6412**
		IMG	5.2585×10^{-7}	0.9999	0.0001	62.7914
	exp	RI	6.8954×10^{-2}	0.2954	0.5676	11.6144
		IMG	2.6718×10^{-4}	0.9168	0.0329	35.7319
	tri	RI	1.2407×10^{-1}	0.1323	0.7122	9.0632
		IMG	2.2469×10^{-5}	0.9956	0.0040	46.4842
10	nvp	RI	6.1564×10^{-2}	**0.4737**	**0.4720**	12.1068
		IMG	1.0395×10^{-6}	0.9998	0.0001	59.8318
	exp	RI	$\mathbf{5.9335 \times 10^{-2}}$	0.3034	0.5311	**12.2669**
		IMG	2.8801×10^{-4}	0.9098	0.0365	35.4059
	tri	RI	8.6142×10^{-2}	0.1301	0.6537	10.6479
		IMG	3.6801×10^{-5}	0.9920	0.0075	44.3414
20	nvp	RI	4.4634×10^{-2}	**0.4030**	0.4500	13.5034
		IMG	9.6994×10^{-6}	0.9985	0.0010	50.1326
	exp	RI	$\mathbf{3.9627 \times 10^{-2}}$	0.3973	**0.3946**	**14.0201**
		IMG	1.8060×10^{-5}	0.9969	0.0027	47.4328
	tri	RI	7.3572×10^{-2}	0.2256	0.5452	11.3328
		IMG	1.5390×10^{-4}	0.9738	0.0256	38.1276

Table 4. Performance metrics for different methods including NVP (nvp), explicit representation (exp), and triplane representations (tri), and data types including RI and predicted images (IMG) on synthetic tissue sample data across various subsample sizes (5, 7, 10, 20). Metrics include MSE, SSIM, LPIPS, and PSNR, providing a comprehensive assessment of reconstruction quality.

Feature Dimension	Metric	MSE ↓	SSIM ↑	LPIPS ↓	PSNR ↑
10	RI	9.34×10^{-4}	0.8743	0.1160	30.2973
	IMG	1.26×10^{-7}	1.0000	0.0000	69.0132
12	RI	$\mathbf{8.46 \times 10^{-4}}$	0.8744	**0.1050**	**30.7278**
	IMG	1.29×10^{-7}	1.0000	0.0000	68.9044
14	RI	9.42×10^{-4}	0.8714	0.1452	30.2575
	IMG	1.91×10^{-7}	1.0000	0.0000	67.1958
16	RI	1.38×10^{-3}	**0.9317**	0.1813	28.6035
	IMG	2.51×10^{-5}	0.9988	0.0026	45.9979
18	RI	1.50×10^{-3}	0.7351	0.2488	28.2328
	IMG	4.92×10^{-7}	1.0000	0.0000	63.0800
20	RI	1.40×10^{-3}	0.7667	0.2210	28.5367
	IMG	3.83×10^{-7}	1.0000	0.0000	64.1675
22	RI	2.82×10^{-3}	0.6420	0.3490	25.5021
	IMG	9.52×10^{-7}	1.0000	0.0000	60.2152
24	RI	1.45×10^{-3}	0.7569	0.2259	28.4001
	IMG	5.99×10^{-7}	1.0000	0.0000	62.2278

Table 5. Metrics for RI and IMG across various test cases with different numbers of features.

Network Layers	Metric	MSE ↓	SSIM ↑	LPIPS ↓	PSNR ↑
2	RI	1.5813×10^{-3}	0.8867	0.2071	28.0100
	IMG	1.5323×10^{-5}	0.9993	0.0017	48.1465
4	RI	8.5690×10^{-4}	0.8781	0.1025	30.6707
	IMG	1.0422×10^{-7}	1.0000	0.0000	69.8203
6	RI	$\mathbf{8.1845 \times 10^{-4}}$	0.8917	0.0918	**30.8701**
	IMG	7.6782×10^{-8}	1.0000	0.0000	71.1474
8	RI	8.1915×10^{-4}	**0.8982**	**0.0898**	30.8664
	IMG	7.2993×10^{-8}	1.0000	0.0000	71.3672
10	RI	1.5813×10^{-3}	0.8867	0.2071	28.0100
	IMG	1.5323×10^{-5}	0.9993	0.0017	48.1465

Table 6. Performance metrics for NVP with varying network layers. The table reports MSE, SSIM, LPIPS, and PSNR for both RI and IMG results, showcasing the impact of network depth on reconstruction quality. The best PSNR for RI is achieved at 6 layers, balancing accuracy and efficiency.

Condition	Metric	MSE ↓	SSIM ↑	LPIPS ↓	PSNR ↑
With self-calibration	RI	$\mathbf{5.71 \times 10^{-3}}$	**0.3911**	**0.6732**	**22.4331**
	IMG	1.51×10^{-3}	0.9117	0.0540	28.2205
Without self-calibration	RI	1.28×10^{-2}	0.2989	0.8297	18.9189
	IMG	1.12×10^{-3}	0.9871	0.0133	29.4935

Table 7. Metrics for RI and IMG results under noise conditions (with and without self-calibration).