MicroDiffusion: Implicit Representation-Guided Diffusion for 3D Reconstruction from Limited 2D Microscopy Projections

Mude Hui[1]* Zihao Wei[2]* Hongru Zhu[3] Fei Xia[4]† Yuyin Zhou[1]†
[1]University of California, Santa Cruz [2]University of Michigan, Ann Arbor
[3]Johns Hopkins University [4]Ecole Normale Supérieure de Paris
muhui@ucsc.edu zihaowei@umich.edu {hongruzhu95, zhouyuyiner}@gmail.com fx43@cornell.edu

Abstract

Volumetric optical microscopy using non-diffracting beams enables rapid imaging of 3D volumes by projecting them axially to 2D images but lacks crucial depth information. Addressing this, we introduce MicroDiffusion, a pioneering tool facilitating high-quality, depth-resolved 3D volume reconstruction from limited 2D projections. While existing Implicit Neural Representation (INR) models often yield incomplete outputs and Denoising Diffusion Probabilistic Models (DDPM) excel at capturing details, our method integrates INR's structural coherence with DDPM's fine-detail enhancement capabilities. We pretrain an INR model to transform 2D axially-projected images into a preliminary 3D volume. This pretrained INR acts as a global prior guiding DDPM's generative process through a linear interpolation between INR outputs and noise inputs. This strategy enriches the diffusion process with structured 3D information, enhancing detail and reducing noise in localized 2D images. By conditioning the diffusion model on the closest 2D projection, MicroDiffusion substantially enhances fidelity in resulting 3D reconstructions, surpassing INR and standard DDPM outputs with unparalleled image quality and structural fidelity. Our code and dataset are available at https://github.com/UCSC-VLAA/MicroDiffusion.

1. Introduction

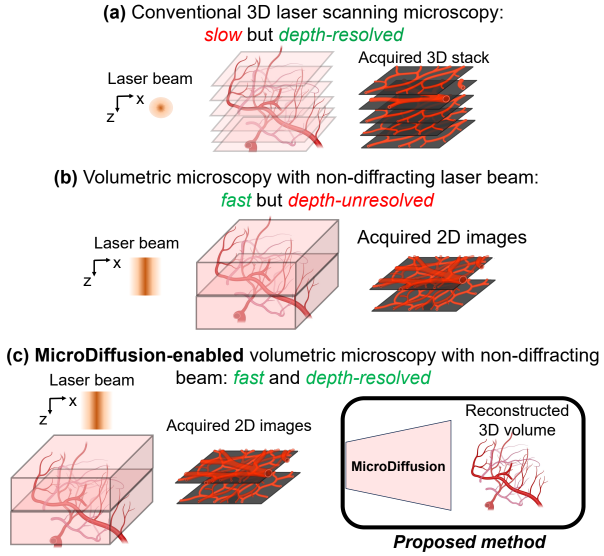

(a) Conventional 3D laser scanning microscopy:
slow but depth-resolved
Laser beam
Acquired 3D stack
z
x
(b) Volumetric microscopy with non-diffracting laser beam:
fast but depth-unresolved
Laser beam
Acquired 2D images
z
x
(c) MicroDiffusion-enabled volumetric microscopy with non-diffracting beam: fast and depth-resolved
Laser beam
z
x
Acquired 2D images
Reconstructed 3D volume
MicroDiffusion
Proposed method

Figure 1. Background and concept of MicroDiffusion-enabled volumetric microscopy. (a) Conventional 3D laser scanning microscopy, while depth-resolvable due to its point-scanning 3D data acquisition scheme, suffers from slow imaging speed. (b) Volumetric microscopy using a non-diffracting laser beam provides fast volumetric imaging by axially projecting 3D volumes onto 2D images but lacks depth information within each acquired 2D image. (c) Our proposed MicroDiffusion model is employed as a digital backend for 3D volumetric reconstruction from 2D projections acquired in (b). MicroDiffusion significantly enhances volumetric imaging performance, providing a synergistic balance between imaging speeds and depth-resolving capabilities.

Volumetric optical imaging has emerged as a pivotal tool in biological and medical domains, enabling precise 3D visualization of intricate structures with unprecedented temporal resolution [13, 46]. Despite its high spatial resolution, the predominant approach in optical microscopy, reliant on 3D laser scanning, suffers from suboptimal temporal resolution due to the slow data acquisition inherent in point-scanning methods (Fig. 1(a)). This limitation not only restricts clinical diagnosis mostly to 2D imaging, potentially compromising diagnostic accuracy [6, 12], but also impedes the observing of dynamic 3D biological processes [14, 18].

Recent advancements using non-diffracting beams have expedited laser scanning microscopy by optically projecting 3D volumes as 2D projections for volumetric imag-

*Denotes equal first author
†Denotes equal senior author

ing [8, 30, 42]. However, this approach sacrifices depth information within each 2D snapshot [3, 43] (see Fig. 1(b)), necessitating the development of tools capable of reconstructing depth for accurate 3D reconstruction from 2D images. In this paper, we aim to reconstruct 3D volumes from limited 2D projections obtained through such volumetric imaging, striving to expedite optical volumetric imaging without compromising depth resolvability of 3D volumes.

Existing 3D reconstruction methods, such as Implicit Neural Representations (INR) [38], offer a comprehensive global view from given 2D microscopy projections by mapping coordinates to a holistic 3D volume using neural networks. However, direct 3D reconstruction using INR often yields globally coherent yet visually blurry outputs, lacking local details. This limitation in spatial resolution may stem from the limited number of 2D images acquired. Conversely, Denoising Diffusion Probabilistic Models (DDPM) [17], especially with the U-Net architecture [32], excel in detailed generative modeling, managing spatial hierarchies, and preserving fine-grained details. Building upon these strengths, we propose MicroDiffusion, a hybrid approach integrating INR's global structural coherence with DDPM's detail enhancement capabilities.

MicroDiffusion encompasses two key designs as shown in Fig. 2: 1) **INR pretraining**, which transforms 2D projections into a preliminary 3D volumetric output, establishing a global structure; and 2) **implicit representation-guided diffusion**, where the pretrained INR acts as a global prior guiding a diffusion model, enhancing details and reducing noise in local 2D projections within the 3D volume. Specifically, MicroDiffusion employs a linear interpolation of INR model output with the noise input, rather than starting from a conventional Gaussian noise baseline, to enrich the diffusion process with structured 3D information. Furthermore, this step enhances image fidelity by conditioning image and positional embeddings extracted from the closest projections. Thus, MicroDiffusion generates 3D reconstructions faithfully representing original optical microscopy images.

Comprehensive experiments on three optical microscopy datasets showcase MicroDiffusion's efficacy. Compared to the baseline INR, it notably enhances reconstruction quality by up to 15.5% in SSIM, 15.2% in PSNR, and 64.7% in DICE on Dendrite dataset, up to 15.0% in SSIM, 3.0% in PSNR, and 0.3% in DICE on Vasculature dataset, and up to 1.8% in SSIM, 0.8% in PSNR, and 4.7% in DICE on Neuron dataset. The resulting 3D stacks demonstrate remarkable resolution, delineating individual dendrites (less than $1\mu m$) and preserving coherent 3D structures—an achievement unattainable by the naive DDPM approach. These tangible outcomes establish MicroDiffusion as a pioneering framework for reconstructing high-quality 3D volumes from 2D projections in volumetric microscopy using non-diffracting beams, reconciling the trade-off between depth information and imaging speed.

2. Related Works

Laser scanning microscopy with non-diffracting beams for volumetric imaging. Laser scanning microscopy has emerged as the gold standard in biomedical imaging. Commonly used in biomedical applications, imaging modalities such as multiphoton microscopy [14], optical coherence microscopy [21], and photoacoustic microscopy [2] all share at least one laser scanning mode. However, a critical challenge in laser scanning microscopy lies in accelerating data acquisition without sacrificing resolution and depth, especially with the most widely used point-scanning methods that scan a tightly focused 3D laser point to collect volumetric data (see Fig. 1(a)).

To address this, optical strategies using non-diffracting laser beams, notably Bessel and Airy beams, have been proposed. These beams generate an elongated and almost uniform axial point spread function. Scanning with non-diffracting beams essentially captures multiple axial layers in a single lateral scan, as opposed to the point-scanning method (see Fig. 1(b)). For instance, they have been utilized for rapid volumetric imaging, enabling the real-time capture of dynamic biological processes [2, 3, 30]. Despite their speed advantages, these strategies often compromise depth information, yielding 2D projections without detailed information on feature depths. To combine the speed benefits of non-diffracting beam scanning methods with the capability to discern depth information, we propose a deep learning model for this inference (see Fig. 1(c)).

Implicit Neural Representations (INR). Implicit neural representations excel at modeling the forms of 3D objects, generating surfaces for 3D scenes, and capturing detailed 3D structures. Pioneering work such as GQN [9] utilizes a generative query network to learn scene representations from multiple perspectives. Building on this foundation, Mildenhall et al. [25] introduce the seminal concept of Neural Radiance Fields (NeRF), which use a multi-layer perceptron to encode 3D scenes for view synthesis. Other works, such as Poly-INR [37], SIREN [34] and LIIF [4], have employed periodic activation functions, significantly enhancing the quality and adaptability of image representations.

Parallel to these advancements, the application of INR in medical imaging has shown remarkable potential. For example, ARSSR [44] and CoIL [40] have adapted NeRF-like methods for super-resolution in medical images. NeRP [35] distinctively combines the inherent image information with the physics of sparse measurements to enhance medical image reconstruction. Cryodrgn [47] and fpm-inr [48] are notable for reconstructing 3D volumes from 2D microscopy images. As a recent advancement, IDM [10] integrates INR and diffusion models by employing INR as the decoder of a diffusion model. In contrast, we leverage INR to generate continuous and interpretable 3D representations used as guidance for a diffusion model.

Diffusion Models. Diffusion models are currently at the forefront of generative model innovation. The Denoising Diffusion Probabilistic Model (DDPM) [16] can incrementally convert Gaussian noise into coherent signals. Subsequent research has expanded on controlling the output of these models, primarily categorized into classifier-guidance [7] and classifier-free guidance [15, 31]. Recent studies demonstrate the versatility of diffusion models in creating content guidance from a variety of sources, including images, text, depth, video, and their combinations [1, 11, 20, 29, 31, 33].

In 3D reconstruction, considerable efforts are made to produce 3D models from text prompts or 2D references [5, 19, 22, 23, 27, 36, 41, 45]. The approach most similar to ours is Magic123 [28], which utilizes a two-stage, coarse-to-fine framework to generate 3D models with reference images. MicroDiffusion differs from Magic123 in several aspects. First, Magic123 employs pretrained knowledge from models like Stable Diffusion [27, 31] or Zero1-to-3 [24] to generate reference views for training Neural Radiance Fields. In contrast, our MicroDiffusion has no such pretrained knowledge, and we must directly train the Implicit Neural Representation (INR) from projection. Second, diffusion in the Magic123's fine stage is applied solely to improve the mesh generated from NeRF, without considering the NeRF's information. In MicroDiffusion, our diffusion model focuses on learning 3D reconstructions, with the INR acting as a source of prior knowledge for global information and actively contributing throughout the training process.

3. Problem Formulation

As depicted in Figure 1(b), a non-diffracting beam creates a uniform point spread function along the axial direction with a limited width, offering an n-fold increase in imaging speed when the optical axial width of the non-diffracting beam is n times that of a conventional point-like axial profile width beam. However, this advantage in volumetric microscopy comes at the cost of depth information, as these beams optically project 3D volumetric information along the axial direction, leading to a lack of depth information in resulting 2D images.

This study aims to develop a model f capable of reconstructing a depth-resolved 3D volume from 2D projections $\{\mathbf{X}_i\}$, obtained using non-diffracting beams. The objective is to achieve a reconstructed 3D volume $\mathbb{M}$ with image quality comparable to traditional point-scanning methods. The 3D stacks that can be acquired from point-scanning methods within the same ith sub-volume are represented as $\mathbb{M}_i = \{m_1^i, m_2^i, \ldots, m_n^i\}$. In non-diffracting volumetric imaging, 2D projections are axially downsampled by a factor of n in $\mathbb{M}_i$, resulting in each projection $\mathbf{X}_i$ being expressed as $\mathbf{X}_i = \frac{1}{n}\sum_{k=1}^{n} m_k^i$. Hence, the problem is to find a model $f : \{\mathbf{X}_i\} \rightarrow \mathbb{M}$ to reconstruct depth-resolved 3D volumes from downsampled 2D projections.

4. Method

In this section, we begin by revisiting key concepts of Implicit Neural Representations (INR) and present our INR design crafted for optical microscopy reconstruction in Sec. 4.1. We then delve into MicroDiffusion, our implicit representation-guided diffusion model in Sec. 4.2.

4.1. Implicit Neural Representation

Revisit INR for 3D Reconstruction. INR methods utilize a function, typically a Multilayer Perceptron (MLP) denoted as f_{inr}, to implicitly represent a 3D field. f_{inr} operates over continuous 3D space and maps coordinates to a predicted property, like intensity or occupancy, formulated as

$$m_{inr} = f_{\text{inr}}(p(z)), \tag{1}$$

where z denotes normalized 3D coordinate within the range $[-1, 1]$ to ensure uniformity across the input space. $p(\cdot)$ denotes the positional encoding that transforms 3D coordinates into a higher-dimensional space, crucial for capturing high-frequency details during reconstruction. m_{inr} represents the property such as the intensity at position z.

Training INR involves a reference dataset, encompassing a set of 2D projections from 3D sub-volumes, as described in Sec. 5.1, with 3D coordinates and corresponding intensities. The training objective is to minimize the reconstruction error between the predicted intensities m_{inr} and the actual data sampled at each coordinate.

INR for Volumetric Microscopy Reconstruction. As shown in Figure 2 (step 1), we sample 3D coordinates uniformly from the 3D volume $\mathbb{M}$, followed by positional encoding and the use of an MLP to map these encodings to voxel density values. For each reference projection $\mathbf{X}_i$, we compute the coordinates for n neighboring slices (defined as 2D projections of neighboring 3D sub-volumes), and concurrently synthesize these slices $\{m_1^i, \ldots, m_n^i\}$ using f_{inr}. Reconstruction loss is measured as the mean squared error (MSE) between the mean of the synthesized slices and the reference projection:

$$L_{\text{mse}} = \sum_{i=1}^{N} \text{MSE}\left(\frac{1}{n}\sum_{k=1}^{n} m_k^i, \mathbf{X}_i\right), \tag{2}$$

where m_k denotes the k-th synthesized slice, and $\mathbf{X}_i$ represents the i-th reference projection.

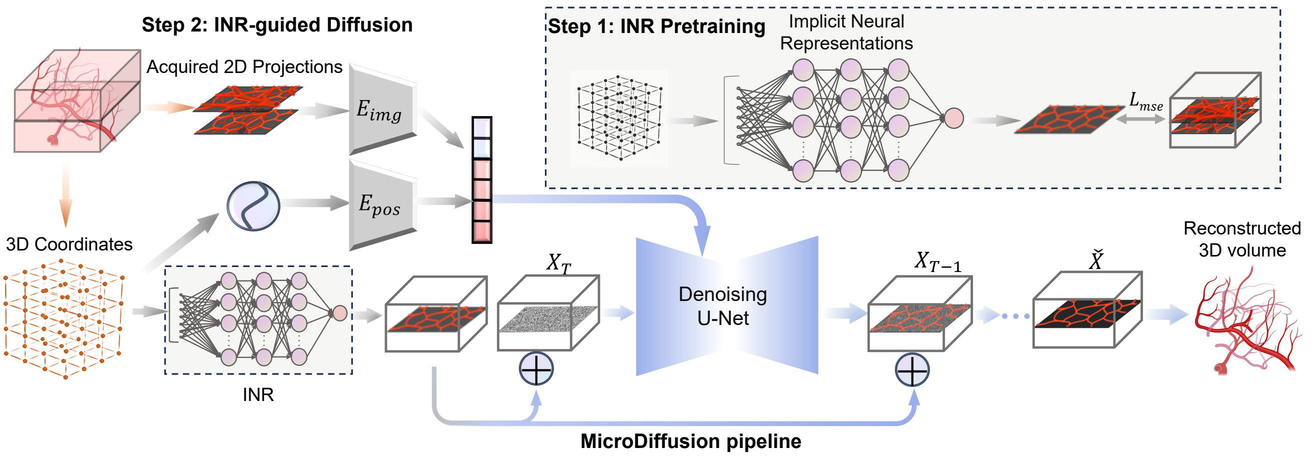


Figure 2. Pipeline of MicroDiffusion. Step 1, we pre-train an INR which provides rough reconstructed images. Step 2, the 2D projections and 3D coordinates are used as the classifier-free guidance of the MicroDiffusion, and the INR output is integrated into the noisy image as guidance during the diffusion process. Detailed information is available at Sec. 4.

INR Neighbouring-based Inference. During training, we simulate the downsampling process triggered by the optical axial projection with a non-diffracting beam, and optimizate the target loss by averaging the output over n coordinates. However, this may introduce a distribution shift when predicting the density solely from its own coordinate. To mitigate this issue, we incorporate information from neighboring slices during inference. When considering a particular 3D coordinate z, the inference result is obtained through a weighted average of n neighboring slices. The weight for the k-th neighboring slice follows a Gaussian distribution:

$$g_k = \frac{1}{\sqrt{2\pi}} e^{-\frac{(\frac{k}{n}-0.5)^2}{2}} \tag{3}$$

While INR reconstruction offers a comprehensive global view, the reconstructed 3D slices suffer from blurriness, artifacts, and lack of fine details (as demonstrated in Figure 3). These issues compromise the spatial resolution and overall reliability of optical microscopy. To address these limitations and ultimately improve reconstruction quality, we introduce a novel approach where we leverage INR as a global prior to guide a diffusion model, enhancing the details and reducing noise in each local 2D slice.

4.2. Implicit Representation-Guided Diffusion

Diffusion Models with Classifier-free Guidance We employ Diffusion Models [17, 39] to reconstruct 3D volumes. As a likelihood model, Diffusion Model can gradually recover the data from Gaussian noise. The forward diffusion process transforms an input X_0 to Gaussian noise $X_T \sim \mathcal{N}(0,1)$ by T iterations, defined as:

$$q(X_t|X_0) = \mathcal{N}(X_t|\sqrt{\bar{\alpha}_t}X_0, (1-\bar{\alpha}_t)I), \tag{4}$$

where X_t represents the data with added noise at time step t. $\bar{\alpha}_t = \prod_{s=0}^{t}(1-\beta_s)$ and β_s represents the noise variance schedule, and $\mathcal{N}$ represents the Gaussian distribution. During training, the neural network $\epsilon_\theta(X_t, t)$ is trained to reconstruct the original data X_0 from the noised data X_t. This is achieved by minimizing ℓ_2 loss between the predicted noise and the actual noise introduced in the data:

$$L(\theta) = \mathbb{E}_{(X,t)}\left[\|\epsilon - \epsilon_\theta(X_t, t)\|^2\right] \tag{5}$$

where t is time step in the forward diffusion process. During the generation process, the neural network $\epsilon_\theta(X_t, t)$ iteratively denoises X_t to achieve high-quality output, θ is the trainable parameters of the model.

Classifier-free Guidance [7, 15] is a method for steering the output generation in Diffusion Models. Diverging from the standard approach of diffusion models, this technique involves training a neural network $\epsilon_\theta(X_t, t, c)$ with an additional conditioning c. The goal is to reconstruct X_0 while incorporating a probability p_{uncond} that $c \leftarrow \emptyset$. The loss function $L(\theta)$ can be written as:

$$L(\theta) = \mathbb{E}_{(X,t)}\left[\|\epsilon - \epsilon_\theta(X_t, t, c)\|^2\right], \tag{6}$$

where $\emptyset$ is the null class. At the generation time, the model uses a guidance scale ω to balance the influence of the conditioning information. This is done by interpolating between the model's predictions with and without the conditioning:

$$\tilde{\epsilon}_t = \epsilon_\theta(X_t, t, c) + \omega \cdot (\epsilon_\theta(X_t, t, c) - \epsilon_\theta(X_t, t)), \tag{7}$$

where $\tilde{\epsilon}_t$ is the noise distribution that model predicts.

Projection and Coordinate Guidance In MicroDiffusion, we use 2D projections and 3D coordinates as conditioning information c in Eq. 6. Projections provide content information of the 3D volume while 3D coordinates provide 3D spatial information.

As depicted in Figure 2, we introduce two distinct encoders: an image encoder, denoted as E_{img}, and a positional encoder, denoted as E_{pos}. E_{img} encodes the current projection $\mathbf{X}_z$, while E_{pos} encodes the 3D coordinate z of the current projection. The ultimate condition information c is formulated as follows:

$$c = E_{img}(\mathbf{X}_z) \oplus E_{pos}(p(z)), \tag{8}$$

where $\mathbf{X}_z$ is the reference projection, $\oplus$ is the concatenate function and $p(\cdot)$ is the coordinate embedding function.

INR Prior Integration To leverage global information and coherent 3D structures in INR to guide the Diffusion Models, we integrate INR outputs as prior knowledge for the diffusion process. Specifically, the INR output m_{inr} is linearly interpolated with the noisy image X_t, which is later to be denoised by Diffusion Models. In MicroDiffusion's training and testing process, for each noisy image X_t at time step t, we perform linear interpolation with the INR output m_{inr} pixel by pixel as

$$X'_t = \gamma m_{\text{inr}} + (1-\gamma)X_t \tag{9}$$

where X'_t is the INR-enhanced image, X_t is the noised image that needs to be denoised by the diffusion model, m_{inr} is the reconstructed output from INR, and γ is the interpolation rate. This approach empowers the diffusion model to directly leverage structural information learned by INR, addressing the learning challenge with a limited number of input 2D projections and enhancing its capacity to generate images with correct 3D structures.

Training and Generation Process MicroDiffusion adopts a conditional U-net [32] similar to that in stable-diffusion [31]. However, in our Denoising U-Net, we remove the cross-attention mechanisms, and add both time condition and conditional feature c at each output of the ResNet block. MicroDiffusion training algorithm is formulated in Algorithm 1, which continues running until convergence. $\oplus$ is the concatenate operation. During training, we first encode 3D coordinates and 2D projections into conditional features c. We then generate the INR prior and the noised data X_t, and linearly interpolate them with an interpolation rate γ. After preparing all model inputs, we follow the equation (6) to update the model parameters.

Algorithm 1 Training function of MicroDiffusion

```
Require: $\mathbf{X}$: 2D projections; $z$: 3D Coordinate; $t$: Time
    step; $p_{\text{uncond}}$ : Probability of being unconditional.
    $m_{inr} = f_{inr}(p(z))$ : INR Inference in Sec. 4.1
    $c = E_{img}(\mathbf{X}) \oplus E_{pos}(p(z))$
    $c \leftarrow \emptyset$ with probability $p_{\text{uncond}}$
    $X_t \leftarrow$sample from $q(X_t \mid X_0)$
    $X'_t = \gamma m_{inr} + (1-\gamma)X_t$
    $L(\theta) = \mathbb{E}_{(X,t)}\left[\|X_0 - \epsilon_\theta(X'_t, t, c)\|^2\right]$
    Take gradient step on $L(\theta)$
```

Algorithm 2 Sampling function of MicroDiffusion

```
Require: $w$: guidance strength; $z$: 3D Coordinate; $\gamma$: in-
    terpolation rate; $T$: Max time step; $\mathbf{X}$: 2D projections.
    $X_T \sim \mathcal{N}(0,1)$
    $c = E_{img}(\mathbf{X}) \oplus E_{pos}(p(z))$
    $m_{inr} = f_{inr}(p(z))$ : INR Inference in Sec. 4.1
    for $t = T$ to $1$ do
        $X'_t = \gamma m_{inr} + (1-\gamma)X_t$
        $\tilde{\epsilon}_t = (1+w)\epsilon_\theta(X'_t, t, c) - w\epsilon_\theta(X'_t, t)$
        $\epsilon_t$ = sample from $\tilde{\epsilon}_t$
        $\tilde{X}_{t-1} = X_t - \epsilon_t$
    end for
    return $X_0$
```

Generation process is outlined in Algorithm 2. Here w is the condition weight controlling whether the model bias more towards conditional or unconditional generation. Similar to the training process, we first prepare all model inputs, and then have the model predict the noise distribution $\tilde{\epsilon}_t$ at the current time-step t. We sample a noise ϵ_t from the noise distribution, subtract it from X_t, and repeat this for T times. We repeat this process for all the coordinates until the algorithm converges.

5. Experiments

5.1. Datasets

We collected experimental data using a conventional multiphoton laser scanning microscope, which has been a gold standard imaging tool for modern biomedical study. This approach is known for creating a three-dimensional, point-like point spread function. By 3D scanning the tightly focused Gaussian beam, we generated high-quality 3D volume stacks. These stacks serve as ground truth datasets for our research problem. Our setup captures 3D volume stacks of various biological structures—such as dendrites, neurons, and vasculature—within the shallow layers of the mouse cortex in a living animal (Fig. 1). These datasets allow us to test our model with varied 2D projected images from diverse 3D biological features of varying densities.

Subsequently, we generated three synthetic datasets. These datasets simulate the case of fast data acquisition using a non-diffracting beam, whose point spread function has a quasi-uniform distribution axially and a predefined width (Fig. 1b). In later experiments, as we will demonstrate, we varied this width to be different multiples (denoted as step length n) of the Gaussian point spread function's axial width used to scan and generate the ground truth datasets. Consequently, the acquired 2D image sequences effectively averaged every n frames along the axial direction, with no spatial overlapping in between. This approach reduced the volume data acquisition time by a factor of n. We then used both the ground truth and the generated synthetic datasets to evaluate our model at different step length n values, focusing on various datasets from the brain. The design of these datasets allows us to directly determine the optimal step length n value, which will inform both future optimal experimental data acquisition and hardware optical design.

5.2. Implementation Details

Depending on the specific imaging modality, the practical axial width of a non-diffracting beam, such as a Bessel beam, can vary from a few times to tens or hundreds of times that of the point-like Gaussian beam used in conventional 3D laser scanning microscopes [3, 13]. We initiate our experiments with a step length n of approximately 6, which corresponds to roughly an order of magnitude in speed-up — an important initial milestone for volumetric imaging. The performances of different reconstruction models were compared at this setting, and subsequently, an ablation study was conducted over the step length value. This study aims to further understand the impact of the step length of n on reconstruction quality, with the goal of identifying the optimal trade-off region between n times speed-up and image reconstruction quality.

For computational efficiency, we downsample all the samples to a resolution of 128 × 128 pixels in the lateral plane. For the pure Implicit Neural Representation (INR) model and the INR encoder, we map the 3D coordinates to a 512-dimensional space using a Gaussian-based embedding technique. The INR model is optimized using the Adam optimizer with a learning rate of 10^{-3} over 5000 epochs, a process that takes approximately 8 hours on an A-100 GPU. Additionally, we employ the AdamW optimizer with a learning rate of 2^{-4} and a weight decay of 10^{-4}. As for the diffusion model, it is trained over 2000 epochs, taking around 4 hours on a single NVIDIA A-100 GPU.

5.3. Baselines

Given the novelty of this task and the absence of existing reference works, we established baseline methods. The initial approach is a straightforward *Interpolation* method, in which the generated structure is created through a uniformly weighted average of the two adjacent projections, weighted according to their distance. And *Interpolation - cubic*, which estimates values by using cubic polynomials between points. This means that each interpolated curve segment is based on the position and slope (derivative) at its endpoints. The last baseline employs a pure INR method, which functions as our prior to the diffusion model.

5.4. Reconstruction Results

5.4.1 Quantitative Results

We evaluate our methods using three metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and the Dice coefficient. PSNR and SSIM are calculated slice-wise along the axial direction, with the mean value across all slices being reported. For the DICE coefficient, we use the OTSU [26] algorithm to determine the threshold for each image to assess the volumetric similarity between the generated 3D structure and the ground truth. As presented in Table 1, our method demonstrates strong performance, successfully capturing the principal structure of the original high-resolution model. Additionally, we observe that the diffusion model's decoder significantly enhances the performance of the pure INR model.

Dataset	Method	SSIM ↑	PSNR ↑	DICE ↑
Dendrite	Interpolation	0.5799	28.78	0.6482
	Interpolation - cubic	0.6511	28.85	0.3973
	INR	0.5837	25.81	0.4589
	Naive Diffusion	0.0297	19.95	0.2869
	Interpolation Diffusion	0.6366	27.02	0.5729
	Interpolation - cubic Diffusion	0.4765	21.31	0.3786
	MicroDiffusion	**0.6742**	**29.74**	**0.7557**
Vasculature	Interpolation	0.3774	20.42	0.5936
	Interpolation - cubic	0.5204	20.52	0.4448
	INR	0.5032	21.69	0.7136
	Naive Diffusion	0.0207	14.81	0.3234
	Interpolation Diffusion	0.4039	19.09	0.4860
	Interpolation - cubic Diffusion	0.2395	16.41	0.2672
	MicroDiffusion	**0.5787**	**22.35**	**0.7158**
Neuron	Interpolation	0.1208	24.12	0.3553
	Interpolation - cubic	0.3265	26.50	0.1116
	INR	0.4759	26.43	0.6403
	Naive Diffusion	0.0210	24.08	0.1468
	Interpolation Diffusion	0.4426	25.35	0.2425
	Interpolation - cubic Diffusion	0.3478	23.79	0.1318
	MicroDiffusion	**0.4845**	**26.66**	**0.6708**

Table 1. Main results of the image reconstruction quality across different datasets with different biological features: vasculature, neurons, and dendrites. For all metrics, higher values indicating better performance as indicated by the arrows.

5.4.2 Qualitative Results

Here, we present the reconstruction results of three methods: pure INR and MicroDiffusion. Part of the slices from the reconstructed 3D stacks are illustrated in Fig. 3, where we randomly selected three slices from the 3D reconstructions generated by naive diffusion, the pure INR reconstruction, our MicroDiffusion and compare with ground truth. From these results, it is evident that the reconstructions obtained via the MicroDiffusion method more closely resemble the ground truth as the density of the biological features increases. This result indicates an encouraging possibility that volumetric imaging with a non-diffracting beam allows not only well-known volumetric imaging of sparse features such as neurons in the cortex [3] but also denser features such as vasculature and even dense dendrites.

5.5. Ablation study

5.5.1 Ablation on conditional feature

How to encode 3D positional information? We initially evaluate two positional encoding methods for MicroDiffu-

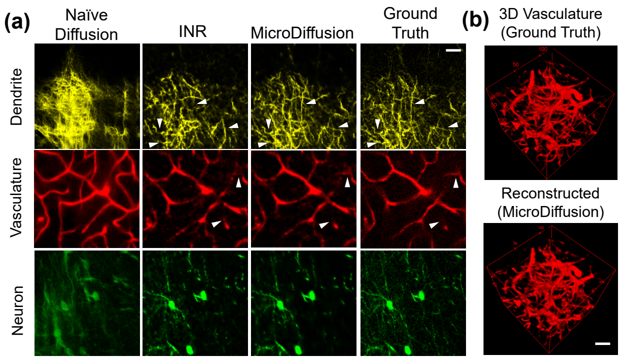


Figure 3. Qualitative results: (a) Comparative visualization of slices from 3D reconstructions with different methods. Observable differences between the INR reconstruction, MicroDiffusion reconstruction, and ground truth are indicated with white arrows. (b) 3D vasculature. Scale bar: 30 μm.

sion: (1) the sine-cosine based encoding as described by NeRF [25], and (2) the Gaussian-based encoding used in NeRP [35]. As shown in Table 2, our experiment results demonstrate that the Gaussian-based encoding yields superior results, particularly in rendering clearer textures. We attribute this improvement to the intrinsic properties of Gaussian embeddings, which affords a more flexible mapping of positions to a higher-dimensional space. This flexibility enhances the subsequent learning process within MicroDiffusion, leading to more detailed and accurate representations.

Method	SSIM↑	PSNR↑	DICE↑
Sin-cos	0.5243	21.03	0.6667
gassian-based(ours)	**0.5787**	**22.35**	**0.7158**

Table 2. Ablation of the positional encoding type on vasculature.

How to add conditional feature? We ablate on the way we incorporate the conditional feature c into the model. *w/o feature guidance* involves setting all conditions to $\emptyset$. *cross-attention* involves adding c using cross-attention to replace the self-attention in Denoising U-net, where the image feature is the query, and the conditional feature c serves as the key and value. Experimental results demonstrate that adding the conditional feature c is effective and leads to the best performance. This is likely because we have only one conditional feature, and in such a case, the cross-attention mechanism may not be effective. Therefore, it is better to directly add the conditional feature to the output of each ResNet block in the Denoising U-net.

5.5.2 Ablation on INR prior

method	SSIM↑	PSNR↑	DICE↑
w/o feature guidance	0.5122	21.27	0.6512
cross-attention	0.5371	21.25	0.6784
addition (ours)	**0.5787**	**22.35**	**0.7158**

Table 3. Ablation results fusion ablation on vasculature

How to generate INR prior? Here, we test three INR prior generation methods as shown in Table 4. *no-neighbouring* means that we only utilize the INR output corresponding to the current 3D coordinate z as the INR prior. *uniformly-mean* means that we use the uniformly averaged output of the INR corresponding to the six frames centered on the current 3D coordinate z. The experimental results demonstrate that our approach performs the best when introducing Neighbouring-based Inference, allowing the model to obtain a more comprehensive 3D INR prior.

method	SSIM↑	PSNR↑	DICE↑
no-neighbouring	0.4315	15.26	0.2025
uniformly-mean	0.4996	17.36	0.4086
Neighbouring-based (ours)	**0.5787**	**22.35**	**0.7158**

Table 4. Ablation of neighbouring based inference on vasculature.

How to add INR prior? We investigate the necessity of the INR prior in this experiment. We trained a naive diffusion model that incorporates the INR prior as the projection introduced in 4.2. We use the image encoder E_{img} to encode the output of INR m_{inr} and concatenate the feature with the other conditions. This allows the model to generate images that resemble true biological features. However, this method performs poorly in acquiring global information, as evidenced by the very low DICE result in Table 5.

Method	SSIM↑	PSNR↑	DICE↑
Naive Diffusion	0.4178	14.97	0.4540
MicroDiffusion	**0.5787**	**22.35**	**0.7158**

Table 5. Ablation of diffusion model INR prior on the vasculature.

5.5.3 Ablation on training method

In our pipeline, we adopt a two-stage training process where the INR is trained initially and then frozen during the MicroDiffusion training. We conducted ablation experiments to explore two alternatives: (1) *joint-training*, where we jointly train INR and the Denoising U-Net from random initialization, and (2) *trainable*, where we unfreeze the INR during the MicroDiffusion training.

We used the same number of epochs for all methods. For *joint-training*, we added the INR loss and applied a decaying weight to balance the training dynamics. As shown in Table 6, our method achieved the best performance. However, *joint-training* proved to be too challenging and adversely affected the MicroDiffusion training process, while *trainable* impaired the ability of INR to provide priors.

Therefore, we chose to train INR first and then train MicroDiffusion with it frozen.

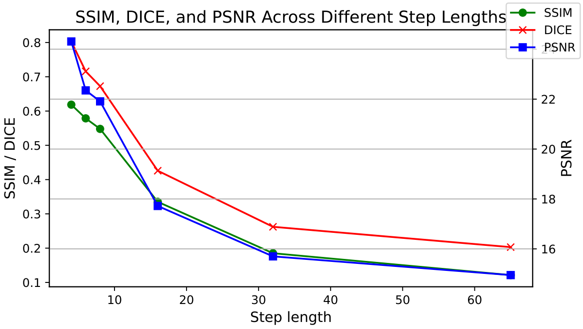
method	SSIM↑	PSNR↑	DICE↑
joint-training	0.5051	21.37	0.6603
trainable	0.5347	21.27	0.6875
freeze (ours)	**0.5787**	**22.35**	**0.7158**

Table 6. Ablation results of training methold on vasculature

5.5.4 Ablation on Different Step length

We investigate the impact of the step length on MicroDiffusion performance. As outlined in our methodology, a larger step size results in faster volumetric imaging but makes reconstruction more challenging. Conversely, smaller step size leads to slower imaging but improved reconstruction. We keep the number of iterations the same and train our model from scratch. Results are presented in Figure 4.

We observed that as the step length increased, all model metrics gradually decreased. To strike a reasonable balance between sampling speed and reconstruction quality, we chose a step length of 6.



(a) SSIM, PSNR and DICE

Figure 4. Performance metrics across different step lengths.

5.5.5 Reconstruction of sparse neuron dataset at various step lengths

A natural question that may arise is whether we can further increase the step length if our features are sparse in space. To address this, Here, we conduct one further experiment aim to assess whether MicroDiffusion can further enhance the speed of volumetric imaging, particularly for samples with sparse spatial distribution. In this context, we have conducted a comparative analysis using the neuron dataset, which is the most sparse case among all the three datasets. The results of this comparison are illustrated in Figure 5. We evaluated the performance of our reconstruction models across a range of step lengths, which correspond to varying degrees of data acquisition speed. Notably, our findings indicate that, in the case of sparse neuron dataset, the step length can be extended to approximately 16 without significantly compromising the quality of the depth-resolved images. This suggests a potential for significant improvements in imaging efficiency without substantial loss in image fidelity for sparser featue of interest.

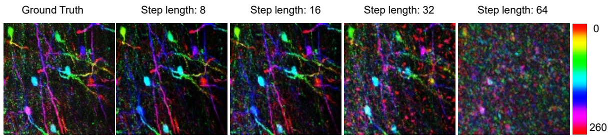


Figure 5. Reconstruction of the depth-resolved sparsely distributed neuron images and depth-resolved volumetric projections with different step lengths.

5.6. Discussion and Future Work

In our experiments, we find that when ground truth data contains Gaussian noise, MicroDiffusion outperforms other methods in noise removal. This demonstrates the potential of MicroDiffusion for denoising 3D volumes acquired by volumetric optical microscopy. In cases where Gaussian noise is intentionally part of the Ground Truth data and should not be removed, it is necessary to investigate how to use other types of noise for MicroDiffusion training.

6. Conclusion

In this paper, we introduce MicroDiffusion, an innovative 3D reconstruction framework that adeptly addresses the challenges of rapid volumetric imaging and the need for depth-rich visualizations in biomedical research. By ingeniously integrating INR with DDPM, MicroDiffusion capitalizes on limited 2D projections to reconstruct high-resolution 3D images, significantly enhancing the capabilities of optical microscopy. Our approach not only accelerates the image acquisition process but also maintains 3D spatial information, allowing for the detailed observation of complex biological structures with minimal data acquisition at high speed. The successful application of MicroDiffusion across various datasets, from densely distributed dendrites to sparsely distributed neurons, underscores its potential as a transformative tool in medical diagnostics and fundamental biomedical research. This work paves the way for designing next-generation volumetric optical microscopy, setting a new benchmark for the integration of machine learning in 3D microscopy volume reconstruction, and opening avenues towards high-speed, high-resolution 3D optical microscopy.

Acknowledgement

This work is partially supported by TPU Research Cloud (TRC) program, and Google Cloud Research Credits program.

References

[1] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 3

[2] Rui Cao, Jingjing Zhao, Lei Li, Lin Du, Yide Zhang, Yilin Luo, Laiming Jiang, Samuel Davis, Qifa Zhou, Adam de la Zerda, et al. Optical-resolution photoacoustic microscopy with a needle-shaped beam. *Nature Photonics*, 17(1):89–95, 2023. 2

[3] Bingying Chen, Xiaoshuai Huang, Dongzhou Gou, Jianzhi Zeng, Guoqing Chen, Meijun Pang, Yanhui Hu, Zhe Zhao, Yunfeng Zhang, Zhuan Zhou, et al. Rapid volumetric imaging with bessel-beam three-photon microscopy. *Biomedical optics express*, 9(4):1992–2000, 2018. 2, 6

[4] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function, 2021. 2

[5] Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tulyakov, Alexander G Schwing, and Liang-Yan Gui. Sdfusion: Multimodal 3d shape completion, reconstruction, and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4456–4465, 2023. 3

[6] Hye Jin Cho, Hoon Jai Chun, Eun Sun Kim, and Bong Rae Cho. Multiphoton microscopy: an introduction to gastroenterologists. *World Journal of Gastroenterology: WJG*, 17 (40):4456, 2011. 1

[7] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021. 3, 4

[8] JJJA Durnin. Exact solutions for nondiffracting beams. i. the scalar theory. *JOSA A*, 4(4):651–654, 1987. 2

[9] S. M. Ali Eslami, Danilo Jimenez Rezende, Frederic Besse, Fabio Viola, Ari S. Morcos, Marta Garnelo, Avraham Ruderman, Andrei A. Rusu, Ivo Danihelka, Karol Gregor, David P. Reichert, Lars Buesing, Theophane Weber, Oriol Vinyals, Dan Rosenbaum, Neil Rabinowitz, Helen King, Chloe Hillier, Matt Botvinick, Daan Wierstra, Koray Kavukcuoglu, and Demis Hassabis. Neural scene representation and rendering. *Science*, 360(6394):1204–1210, 2018. 2

[10] Sicheng Gao, Xuhui Liu, Bohan Zeng, Sheng Xu, Yanjing Li, Xiaoyan Luo, Jianzhuang Liu, Xiantong Zhen, and Baochang Zhang. Implicit diffusion models for continuous super-resolution, 2023. 2

[11] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all, 2023. 3

[12] Jan Goedeke, Peter Schreiber, Larissa Seidmann, Geling Li, Jérôme Birkenstock, Frank Simon, Jochem König, and Oliver J Muensterer. Multiphoton microscopy in the diagnostic assessment of pediatric solid tissue in comparison to conventional histopathology: results of the first international online interobserver trial. *Cancer management and research*, pages 3655–3667, 2019. 1

[13] Sven Gottschalk, Oleksiy Degtyaruk, Benedict Mc Larney, Johannes Rebling, Magdalena Anastasia Hutter, Xosé Luís Deán-Ben, Shy Shoham, and Daniel Razansky. Rapid volumetric optoacoustic imaging of neural dynamics across the mouse brain. *Nature biomedical engineering*, 3(5):392–401, 2019. 1, 6

[14] Fritjof Helmchen and Winfried Denk. Deep tissue two-photon microscopy. *Nature methods*, 2(12):932–940, 2005. 1, 2

[15] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. 3, 4

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pages 6840–6851. Curran Associates, Inc., 2020. 3

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 4

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

[19] Animesh Karnewar, Andrea Vedaldi, David Novotny, and Niloy Mitra. Holodiffusion: Training a 3d diffusion model using 2d images, 2023. 3

[20] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6007–6017, 2023. 3

[21] Kye-Sung Lee and Jannick P Rolland. Bessel beam spectral-domain high-resolution optical coherence tomography with micro-optic axicon providing extended focusing range. *Optics letters*, 33(15):1696–1698, 2008. 2

[22] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 300–309, 2023. 3

[23] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization, 2023. 3

[24] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object, 2023. 3

[25] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 2, 7

[26] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66, 1979. 6

[27] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv*, 2022. 3

[28] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, and Bernard

Ghanem. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors, 2023. 3

[29] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021. 3

[30] Cristina Rodríguez, Yajie Liang, Rongwen Lu, and Na Ji. Three-photon fluorescence microscopy with an axially elongated bessel focus. *Optics letters*, 43(8):1914–1917, 2018. 2

[31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3, 5

[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 2, 5

[33] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement, 2021. 3

[34] Changwon Seo, Kyeong-Joong Jeong, Sungsu Lim, and Won-Yong Shin. Siren: Sign-aware recommendation using graph neural networks, 2022. 2

[35] Liyue Shen, John Pauly, and Lei Xing. Nerp: Implicit neural representation learning with prior embedding for sparsely sampled image reconstruction, 2023. 2, 7

[36] J Ryan Shue, Eric Ryan Chan, Ryan Po, Zachary Ankner, Jiajun Wu, and Gordon Wetzstein. 3d neural field generation using triplane diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20875–20886, 2023. 3

[37] Rajhans Singh, Ankita Shukla, and Pavan Turaga. Polynomial implicit neural representations for large diverse datasets, 2023. 2

[38] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in neural information processing systems*, 33:7462–7473, 2020. 2

[39] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 4

[40] Yu Sun, Jiaming Liu, Mingyang Xie, Brendt Wohlberg, and Ulugbek S. Kamilov. Coil: Coordinate-based internal learning for imaging inverse problems, 2021. 2

[41] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Viewset diffusion: (0-)image-conditioned 3d generative models from 2d data, 2023. 3

[42] Xiao-Jie Tan, Cihang Kong, Yu-Xuan Ren, Cora SW Lai, Kevin K Tsia, and Kenneth KY Wong. Volumetric two-photon microscopy with a non-diffracting airy beam. *Optics Letters*, 44(2):391–394, 2019. 2

[43] Andres Flores Valle and Johannes D Seelig. Two-photon bessel beam tomography for fast volume imaging. *Optics express*, 27(9):12147–12162, 2019. 2

[44] Qing Wu, Yuwei Li, Yawen Sun, Yan Zhou, Hongjiang Wei, Jingyi Yu, and Yuyao Zhang. An arbitrary scale super-resolution approach for 3d MR images via implicit neural representation. *IEEE Journal of Biomedical and Health Informatics*, 27(2):1004–1015, 2023. 2

[45] Jiale Xu, Xintao Wang, Weihao Cheng, Yan-Pei Cao, Ying Shan, Xiaohu Qie, and Shenghua Gao. Dream3d: Zero-shot text-to-3d synthesis using 3d shape prior and text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20908–20918, 2023. 3

[46] Seok H Yun, Guillermo J Tearney, Benjamin J Vakoc, Milen Shishkov, Wang Y Oh, Adrien E Desjardins, Melissa J Suter, Raymond C Chan, John A Evans, Ik-Kyung Jang, et al. Comprehensive volumetric optical microscopy in vivo. *Nature medicine*, 12(12):1429–1433, 2006. 1

[47] Ellen D. Zhong, Tristan Bepler, Joseph H. Davis, and Bonnie Berger. Reconstructing continuous distributions of 3d protein structure from cryo-em images, 2020. 2

[48] Haowen Zhou, Brandon Y. Feng, Haiyun Guo, Siyu Lin, Mingshu Liang, Christopher A. Metzler, and Changhuei Yang. Fpm-inr: Fourier ptychographic microscopy image stack reconstruction using implicit neural representations, 2023. 2