Clean Implicit 3D Structure from Noisy 2D STEM Images

Hannah Kniesel, Ulm University
Timo Ropinski, Ulm University
Tim Bergner, Ulm University
Kavitha Shaga Devan, Ulm University
Clarissa Read, Ulm University
Paul Walther, Ulm University
Tobias Ritschel, University College London
Pedro Hermosilla, Ulm University

Abstract

Scanning Transmission Electron Microscopes (STEMs) acquire 2D images of a 3D sample on the scale of individual cell components. Unfortunately, these 2D images can be too noisy to be fused into a useful 3D structure and facilitating good denoisers is challenging due to the lack of clean-noisy pairs. Additionally, representing detailed 3D structure can be difficult even for clean data when using regular 3D grids. Addressing these two limitations, we suggest a differentiable image formation model for STEM, allowing to learn a joint model of 2D sensor noise in STEM together with an implicit 3D model. We show, that the combination of these models are able to successfully disentangle 3D signal and noise without supervision and outperform at the same time several baselines on synthetic and real data.

1. Introduction

STEMs enable the acquisition of 3D samples from 2D images on the scale of cellular components [16, 20]. This allows for addressing many important tasks in biology, that rely on the spatial organization inside cells [31, 47].

In STEMs, the amount of electrons used to probe a sample needs to be low, in order to prevent sample damage as well as to keep acquisition times at bay [28]. This, unfortunately, leads to 2D images that can be noisy. While many sophisticated image denoisers exist, fusing noisy 2D into consistent 3D structure poses its own challenge. Many forms of fusing 2D information into 3D [54] assume that the same world point has the same properties in all its 2D projections. In the presence of noise, this assumption does not hold. While this might be negligible in many instances of fusing photographic-domain imagery taken under normal lighting conditions, the noise in the electron domain is much more intricate, *i.e.*, it is, first, strong, and, second, does not follow a simple Gaussian model. Thus, our first contribution is to model this 2D noise for STEMs using Normalizing

Corresponding author email address: hannah.kniesel@uni-ulm.de

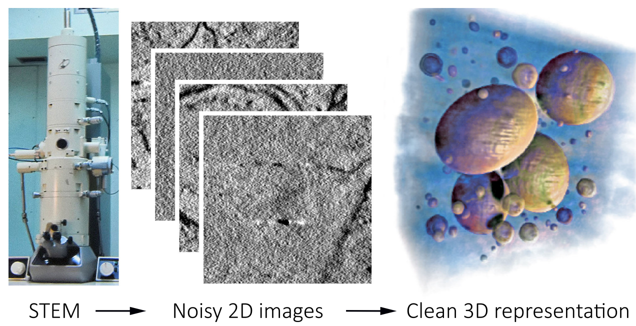


Figure 1. Our approach enables learning of clean 3D structure from very noisy 2D sensor readings as produces by STEM.

Flows [32, 50] in an unsupervised setup.

To establish a link between 2D observations and a 3D model, a range of recent methods employ differentiable volume rendering [29, 57], which allow for changing 3D information so that when rendered, it matches some input. For these techniques, besides the difficulty of handling noise, representing detailed 3D structures can be challenging even for clean data, when using regular 3D grids. Fortunately, implicit models like occupancy fields [15, 42, 52] or Neural Radiance Fields (NeRF) [44] have recently shown great potential to represent 3D structures from photographs. These methods do not rely on a regular 3D grid, as they learn a 3D function to represents the shape itself. The loss of this learning requires to project the 3D representation to 2D images to be compared to the 2D observations. Our second contribution is to unleash these methods for STEM reconstruction, by deriving the projection for STEM and a Maximum-likelihood Estimation (MLE) loss to compare the outcome to noisy observations. This does not introduce the blur arising from the L1 or L2 loss commonly used in NeRF.

We will apply both our contributions jointly and show that this combination can successfully disentangle 3D signal and noise without supervision, and outperform at the same time both standard reconstruction algorithms and all variants of our setup where the noise is not modeled. We make all our data, code and trained networks available [1].

[1] https://github.com/HannahKniesel/Implicit-Electron-Tomography.git

2. Previous Work

Inverse problems: 3D reconstruction from STEM images is an instance of an inverse problem. Inverse problems aim to recover a signal from indirect measurements where the process to obtain such measurements is known. This is modeled using a forward operator F transforming the signal x, which we aim to recover, into the observations, $o = F(x)$. Additionally, these observations are usually affected by noise.

A set of well established algorithms to solve these problems are Back Projection algorithms [18, 49]. However, such methods greatly suffer from artifacts when the number of observations is limited, which is usually the case in STEM. Other algorithms have tried to solve the problem of limited observations using iterative algorithms [7, 23] with regularizers to enforce continuity on the reconstructed data such as L2, L1, or Total Variation (TV). These problems have been studied on different fields, such as reconstruction from electron microscope images [5, 39, 46, 58], X-Ray computed tomography (CT) [6, 10, 17, 62], or visible light tomography [61]. However, 3D reconstruction from STEM poses additional challenges such as the low number of observations, the *missing wedge problem* due to the limited angle range used, and the large image sizes which translate to large memory requirements. For a more thorough review of these 3D reconstruction methods for different electron microscope modalities we refer the reader to surveys by Sorzano et al. [53] and Frank [20].

In the last years, a new set of data-driven methods have been proposed to solve inverse problems for CT and MRI data. These methods have addressed the problem by pre-processing the observations [8], post-processing the reconstruction [37, 59], learning the reconstruction process [26, 66], by using iterative approaches [6], or by overfitting a neural network to a single reconstruction [10, 62]. Unfortunately, these methods have assumed a simplified noise model using a Poisson-Gaussian distribution [2].

In the field of Electron Microscopy (EM), deep learning has been recently applied to single-particle reconstruction from Cryo-EM images. Gupta et al. [24, 25] proposed a 3D reconstruction using a volumetric representation and trained it using a GAN objective. Zhong et al. [64, 65] recently proposed a method to represented the 3D reconstruction in Hartley space using a neural network and optimized the reconstruction and the pose information of each image together. However, those methods rely on a large number of images covering all possible view directions for the reconstruction, and also assume a simple noise model. Recently, a new deep learning approach has been suggested to improve reconstruction based on STEM images [1]. However, this method is composed of a denoiser network and a super-resolution module to improve the reconstructed volume obtained with standard algorithms. In this work instead, we represent the 3D reconstruction with neural networks and learn it with a limited number of observations while simultaneously modelling the observed noise in an end-to-end framework not requiring supervision by clean images.

Implicit reconstruction: Recently, representing a 3D scene implicitly using neural network has gained a lot of attention [41, 43, 45]. These neural networks receive as input 3D coordinates of a point in space and output the signed distance to the surface of the object. This concept was later extended using localized neural representations to only store information in the occupied parts of the scene [21, 22, 55, 56]. Further work extended these ideas and use such representations for 3D scene reconstruction from multiple images [44]. They proposed a neural network to encode not only the occupancy in the scene, but also the radiance at each 3D location and output direction. Thanks to transforming input coordinates and view direction using positional encoding, they were able to achieve high quality reconstructions. Several works have followed up and proposed different improvements [11, 27, 60]. Recent work have also used similar ideas for 3D reconstruction from different image modalities [9].

Noise modelling: Image noise is an undesired perturbation of the measured intensity of a pixel generated by errors in the image acquisition process. The most common noise models used in image denoising algorithms assume an additive white Gaussian noise [65], a Poisson-Gaussian model [2, 19, 63], or a Gaussian distribution with pixel dependent variance [36]. Recent works have also suggested to train neural networks to denoise images in an unsupervised setting by imposing certain restrictions on the type of noise model [12, 13, 14, 34, 35]. Unfortunately, these simple models and strict constrains are not able to cover certain noise sources arising during the process of image generation in EM data [20]. In another line of research, Abdelhamed et al. [3] have used Normalizing Flows to model the noise distribution from data in a supervised setting, without making any assumption of the underlying noise model. In our work, we use Normalizing Flows to model the noise but we learn it in an unsupervised fashion thanks to the spatial constrains introduced by the 3D reconstruction process.

3. Our Approach

3.1. Overview

We will first present an overview of our approach (Sec. 3.1) before describing the STEM image formation (Sec. 3.2), an implicit 3D representation of the result (Sec. 3.3) and a noise model (Sec. 3.4) for STEM. We conclude on details of the loss used (Sec. 3.5).

Our system is modeled after NeRF but with two important generalizations (Fig. 2): First, where NeRF is modelling an emission-absorption model for photons [44], we consider a model for electrons. Second, we do not map from

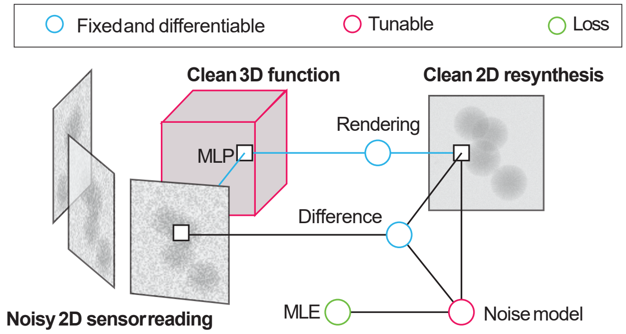


Figure 2. Overview of our approach. Given a set of n observations, $\hat{E}_i$, we optimze the parameters of two tunable mappings. The first is an implicit 3D model (MLP, pink box) that represents the geometry as a continuous implicit density field. The second is a mapping from clean values of accumulate density, *i.e.* opacity, to noisy sensor readings (Noise model, pink circle). 3D and 2D are linked by a fixed image formation model (blue circle) and compared using a loss capable to compare distributions (MLE, green).

a 3D solution to a single 2D image, but to a distribution of images, including the noise. This prevents converging to the mean of the noise, which is not the correct value. We will explain both parts in the next sections.

3.2. Image formation model

Our image formation is comprised of a ray-marching variant suitable for noisy and out-of-focus electron beams.

Acquisition setup: The image acquisition process in STEM uses an electron beam that is focused at a point within the sample as it is illustrated in Fig. 3, a. Electrons that pass through the specimen from the top are then captured by the detector. This process is repeated for all pixels in the image by displacing the sample by a certain distance, p_s as seen in Fig. 3, b. Once the image is complete, the sample is tilted by α degrees and the spatial scan is repeated until the desired number of images is captured. Common existing hardware allows for a tilt of up to ± 72 degrees [20].

Raymarching for electrons: We here adopt the pinhole emission-absorption model now often used in differentiable volume rendering [29, 44] to electron beams.

In an absorption-only model [40], the fraction of electrons lost

$$\frac{\mathrm{d}E(\mathbf{r}(t))}{\mathrm{d}t} = -\sigma(\mathbf{r}(t))E(\mathbf{r}(t)) \tag{1}$$

for an infinitesimal step $\mathrm{d}t$ at distance t along a ray $\mathbf{r}(\mathbf{x}+t\cdot\omega)$ from position $\mathbf{x}$ in direction ω is proportional to the density $\sigma(\mathbf{r}(t))$ of the medium at that position along the ray. Electrons, technically, are not absorbed but scattered into many different forms of secondary emissions, we ignore here as the detector is small compared to the distance to the sample, and almost all scattered electrons will not arrive at the detector. Also, all electrons have the same energy in STEM, and hence density does not depend on what would be wavelength or color for photons in the optical regime. This equation has the solution [40]

$$E(\mathbf{r}) = \exp\left(-\int_0^t \sigma(\mathbf{r}(t))\mathrm{d}t\right). \tag{2}$$

The inner integral can be solved by numeric quadrature, *i.e.* as a sum or using Monte Carlo estimation.

Defocus: Above considerations assume an infinitely small ray, while in reality, the contribution to the readings of a detector pixel is the confound effect of a bundle of rays. Hence the electron beam is not a double-cone but a double-wedge as seen in Fig. 3, c. A system is in focus, if the ratio $p_\mathrm{s}/d_\mathrm{s}$ between the pixel distance and the width of the electron beam is larger than 1. The example in Fig. 3, b/c is in focus, as $p_\mathrm{s} > d_\mathrm{s}$ and so is the setup used in our results with $p_\mathrm{s}/d_\mathrm{s} = 1.86$ for a tilt angle of $\alpha = 0$.

If the sample is tilted however (Fig. 3, d), locations at distance from the tilt axis move out of the focal plane, resulting in out-of-focus blur. While the hardware accounts for the angle of the tilt α, there is always a small error in angle, resulting in a residual angle β. Not accounting for this aspect will result in a system which learns the blur or a mix of blurry and sharp observations when seeing one world point under different (residual) tilt angles.

The full out-of-focus image formation could be solved via Monte-Carlo integrating not only the path integral from Eq. 2, but also an integral over an area of all sensor locations A and a set of directions Ω for a pixel P

$$E(P) = \int_A \int_\Omega \exp\left(-\int_0^t \sigma(\mathbf{r}(\mathbf{x},t))\mathrm{d}t\right) \mathrm{d}\omega \mathrm{d}\mathbf{x}. \tag{3}$$

To solve this integration problem effectively, we seek inspiration from Computer Graphics approaches for realistic simulation of lenses [48], in particular screen space methods [51]. These represent the 5D double integral in Eq. 3 as a 2D integral in image space instead. This integral then becomes a spatially-varying convolution of per-pixel radiance $E(\mathbf{r})$ with a Point-spread Function (PSF) of the optical system. The action can be described by the convolution

$$E(P) = \kappa(\beta, d) * E(\mathbf{r}) \tag{4}$$

$$\kappa(\mathbf{x})(\alpha, d) = \exp(-||\mathbf{x}|| \cdot \tan(\alpha) \cdot d) \tag{5}$$

with a single blur kernel κ that depends on the tilt angle β and the image-space distance d from the tilt axis. While the true PSF of STEM might have a different shape, the qualitative low-pass is reproduced by this Gaussian which is fast to execute and well-differentiable.

3.3. 3D Representation

In EM, density distribution σ in 3D space is classically modeled as a discrete grid. We follow the recent trend

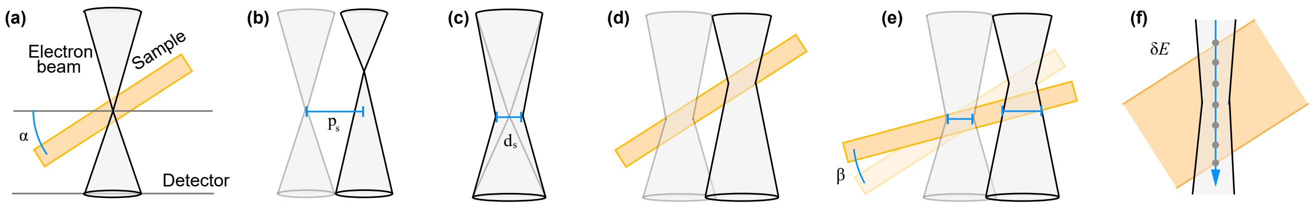


Figure 3. STEM imaging process: **(a)** Ideal double-cone set up of a single pixel. **(b)** Pixel distance and refocus. **(c)** Non-ideal double-wedge setup and circle of confusion. **(d)** Non-ideal tilt and resulting variation in the circle of confusion. **(e)** Electron transport in a slab.

[15, 42, 44] to represent such 3D fields as an implicit function instead. We use an Multi-layer perceptron (MLP) σ_θ that maps position to density. For details on the MLP architecture, we refer the reader to the supplementary material.

Notation: Shorthand, we will drop the dependency of $E(\mathbf{r})$ on $\mathbf{r}$ and will write E for the radiance of some ray. On occasion, we will further write E^θ, to denote the radiance resulting from tracing a certain ray through an implicit field defined by an MLP chosing parameters θ.

3.4. Noise model

Unfortunately, we can only measure a noisy estimate of the true number of electrons per unit space, time and solid angle. This is, both because of the quantized nature of the electron beam, resulting in Poisson-like noise, and due to other sources of noise, in particular from the requirement to turn the electron beam into light to be read by a photo-sensitive A/D conversion.

Instead of observing E, we have to deal with samples from $p(\bar{E}|E)$, stating the probability density of observing $\bar{E}$ when the true value is E.

Were we given pairs of clean and noisy values $\bar{E}$ and E it would be simple to train a generative noise model. In the case of STEM however, it is difficult to acquire pairs of clean and noisy sensor readings as the sensing process itself changes the sample while at the same time depending on the sample.

As a remedy, we train the noise model jointly with the 3D density field itself. The key insight here is that there is multiple noisy observations of the same clean density field, but under different rays.

We use Normalizing Flows [32, 50], as this can both: compute the density $p(\bar{E}|E)$, the probability of $\bar{E}$ given E (as required by our loss we describe next) and generate samples $q(\xi|E) \sim p$ where ξ is a random number. We will write shorthand q^ϕ for an instance of the noise model with tunable parameters ϕ.

In particular, we use eight 1D radial flow layers [50] which transforms a Gaussian distribution into our desired $p(\bar{E}|E)$ distribution. These layers apply radial contraction and expansion around a reference point and are defined as:

$$f(z) = x + \frac{\beta(z - z_0)}{\alpha + |z - z_0|}$$

where the learned parameters are $z_0 \in \mathbb{R}$, $\alpha \in \mathbb{R}^+$, and $\beta \in \mathbb{R}$. To condition the noise distribution to the real radiance, E, and therefore allow for modeling signal-dependent noise, in the last four layers of our Normalizing Flow model, the learned parameters $\{z_0, \alpha, \beta\}$ are predicted by a small MLP which takes as input E.

Recent work has suggested to use more complex Normalizing Flow models to learn a noise distribution from noise–clean pairs [3]. Such models use CNN layers to condition the learned distribution on a region of the clean image. Unfortunately, allowing the noise model to inspect the image could allow the model to not only learn the noise distribution but also to fix artifacts and missing details that the 3D reconstruction was not able to recover. By conditioning the model on a single pixel, our reconstruction framework is able to separate the 3D signal from the noise.

Implementation: In this section we describe the target probability density function as $p(\bar{E}|E)$. Instead, we learn the distribution of differences between the true and the observed radiance, $p(\bar{E}-E|E)$. This objective is equivalent to the one described in this section. However, this distribution allows for the gradients to propagate not only through the flows' conditioned layers, but also directly from the loss.

3.5. Loss

We are looking for a scalar density field $\sigma^\theta(\mathbf{x}) \in \mathbb{R}^3 \to \mathbb{R}$ as well as a noise model $q^\phi(\xi) \in \mathbb{R} \to \mathbb{R}$ with tunable parameters θ and ϕ, to explain the observed opacities $\hat{E}_i$. We can compute the clean solution for pixels in images as we know their relative orientations, camera geometry and hence, the ray $\mathbf{r}_i$ of every pixel.

Clean case: With access to clean observations, we could minimize the empirical risk

$$\arg\min_\theta \mathbb{E}_i[\mathcal{L}(E_i^\theta, \hat{E}_i)].$$

Recall, that clean samples E do not exist and we have to work with noisy samples $\bar{E}$, leading to the following thought experiment:

Consider the case of a volume of constant density θ and an optimization to find this density given multiple noisy observations of that volume $\hat{E}_i$. If we were to minimize this under the $\mathcal{L}_2$ loss, it was to produce the mean of all density solutions explaining all observation. Under an $\mathcal{L}_1$, it would converge to the median of all density solutions. Importantly, neither mean or median of the distribution of solutions is the ground truth value θ of complex STEM noise.

Noisy case: The key to make this work, is to match the entire distribution of noisy observations to a generative model producing a distribution of radiances. The only combination to explain those distribution pairs is a parameter pair θ/ϕ: a clean 3D volume that, after ray-marching and defocus blur, and after adding synthetic noise, produces the observation distribution. Hence, we minimize

$$\arg\min_{\theta,\phi} \mathbb{E}_{i,\xi}[\mathcal{L}_{\text{MLE}}(p^{\phi}(\hat{E}_i - E_i^{\theta}|E_i^{\theta}))].$$

Note, that the distribution loss is defined on the difference of noisy observations in respect to the clean ones. This allows for learning the correct solution up to an additive constant. As most applications are not concerned with getting the exact absolute value (such as a photo does not tell the absolute radiance unless we know exposure, aperture and ISO), this might be an acceptable limitation in most – but not all– applications. Noteworthy, this is not a limitation of knowing the exact physical parameters of the STEM but a core limitation of our approach to denoising.

4. Results

We present results of our approach in different datasets, for different methods according to different metrics which we will all explain next.

Data: We consider a SYNTHETIC and a REAL data set. The main motivation is, that to our knowledge no ground truth data for a quantitative evaluation of real STEM acquisitions is available. Recall, that training proceeds from scratch for every data set. For every data set we have a certain number of noisy 2D images available, that is split into test, train and validation.

The SYNTHETIC data set is comprised of random arrangements of ellipsoidal shells of a random density and a density model of the ZIKV (*i.e.*, Zika) virion at 15Å by Long et al. [38]. As we know the clean analytic solution, this data set can be used for quantitative evaluation. As it resembles the 3D structure of cells, it allows for qualitative evaluation, too. We assume a 79-image tilt series ranging from -59.5° to 59° with 1.5° steps. Additionally, we generate 14 projections for validation, and 20 for testing. All projections are rendered in 1000×1000 pixels. To add noise to the simulations we train a Normalizing Flow to match the noise distribution of STEM images. We acquire the training data of the Normalizing Flow network by real STEM data. During data acquisition, for each tilt step, two images are taken: One image with short exposure time and one image with long exposure time, resulting in noisy and less noisy image pairs. We make sure the image pairs are aligned pixel-precise, by using the SIFT algorithm of the ImageJ toolbox [4]. To retrieve an accurate alignment, we

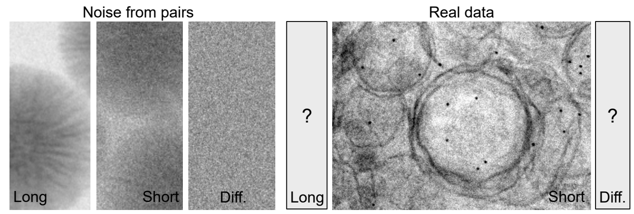


Figure 4. To generate noisy synthetic data for evaluation, we obtain noise samples with the difference between aligned short and long exposure STEM images of a set of nanoparticles (**left**). We also evaluate on real STEM images of cells infected with SARS-CoV-2 where only short exposed images are available (**right**).

image simple structures of nanoparticles. Then we retrieve the difference image which is now containing only the noise of the short exposure STEM. See Fig. 4 for an illustration of this process. We train the Normalizing Flow by conditioning it on the less noisy tilt series to then match the noise distribution of the noisy tilt series. By sampling from the trained Normalizing Flow model, we are able to generate a pair of tilt series, with clean and noisy projections.

The REAL images contain cells infected with SARS-CoV-2 using short exposure times. The tilt series ranges from -72° to 72°, with a tilt step of 1.5° at a resolution of 900×900 pixels. Before reconstruction, we align the raw tilt series with the IMOD software by Kremer et al. [33], using fiducials. For validation we use a projection of the training data. Without a way to capture ground truth for such data, they are used only for qualitative evaluation.

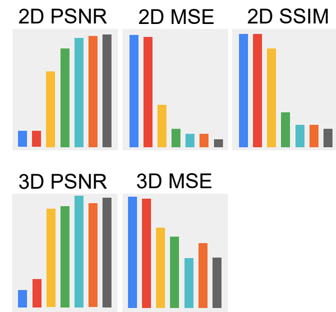
Methods: Besides `Ours`, we consider commonly used reconstruction algorithms for STEM images and several variants of our method. The first methods are the weighted back-projection (`WBP`) and the Simultaneous Iterative Reconstruction Technique (SIRT) method (`SIRT`) implemented in the software package IMOD [33], a state-of-the-art industry solution to tomographic reconstruction problems. Next, we explore using our implicit 3D representation but training directly under and $\mathcal{L}_2$-loss, as done in NeRF, for either the noisy data directly (`L2Noisy`) or the data denoised in 2D with a common denoiser, BM3D (`L2Den`). We study a supervised variant of our approach, which assumes the knowledge of the noise model (`OursSup`). Lastly, `L2Clean` is the same setup as `L2Noisy`, but trained on the clean projections instead. This is an upper bound what the implicit 3D reconstruction can achieve for this data if no noise is present in the observations.

Metric: We consider different metrics on different forms of the results: The full 3D volume and random 2D projections. For 2D we can apply DSSIM, Peak Signal-to-noise Ratio (PSNR) and Mean Squared Error (MSE), in 3D PSNR and MSE. The full 3D volume is discretized to 1000×1000×1000, images are rendered in 1000×1000.

Training Details: For training of the MLP we use an ADAM optimizer with a learning rate of 5^{-5}. For training

Table 1. Main quantitative results of different methods (rows), reconstructing a known ground truth volume according to different metrics (columns). The best method across all methods without access to clean ground truth are shown in **bold**.

Method	Loss	Imp.	Clean.	2D			3D	
				PSNR	MSE	DSSIM	PSNR	MSE
WBP	$\mathcal{L}_2$	×	×	2.97	50.901	4.904	7.62	17.299
SIRT	$\mathcal{L}_2$	×	×	3.27	47.622	4.832	9.13	12.223
L2Noisy	$\mathcal{L}_2$	✓	×	13.86	1.885	4.271	19.73	1.064
L2Den	$\mathcal{L}_2$	✓	×	18.15	0.849	1.544	20.25	0.944
Ours	$\mathcal{L}_{\text{MLE}}$	✓	×	**19.93**	**0.645**	**1.020**	**21.75**	**0.669**
OursSup	$\mathcal{L}_{\text{MLE}}$	✓	×	20.07	0.636	0.991	20.64	0.864
L2Clean	$\mathcal{L}_2$	✓	✓	20.73	0.393	0.840	21.60	0.691



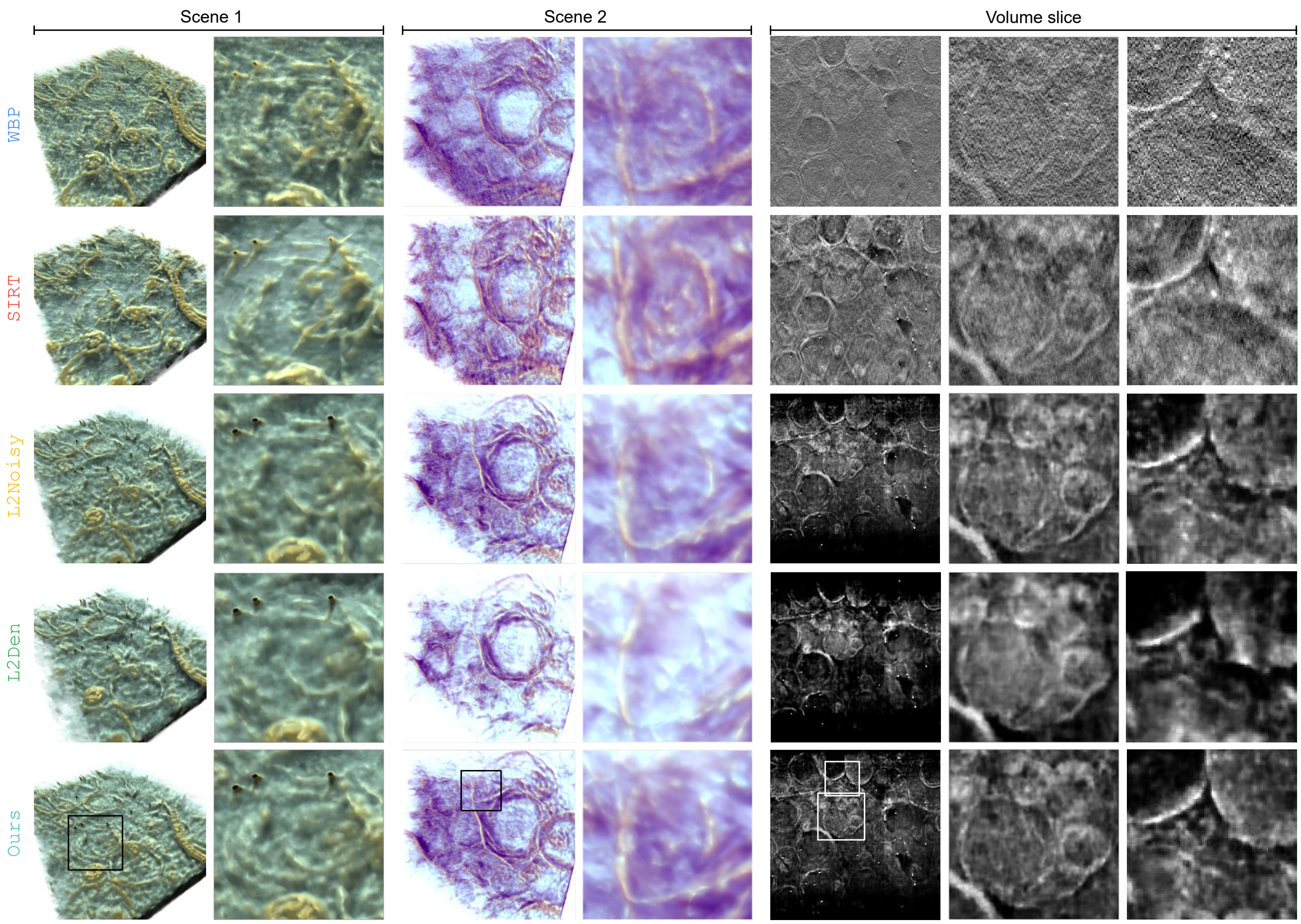


Figure 5. Qualitative results of different methods **(row)** for different slices **(column)** of the REAL data set. Trends of the SYNTHETIC data can be revisited on the REAL data, even though differences in the outcome of the learned approaches are less noticeable. Again, in the presence of noise, WBP and SIRT cannot produce useful output. Training after denoising (L2Den) suppresses small details and results in an overall smoother reconstruction. But differences on L2Noisy and Ours are hard to evaluate, specially without the knowledge of GT.

the Normalizing Flow we found that using the SGD optimizer worked the best. For the REAL data set we use 512 neurons in the hidden layers of the 3D reconstruction module, while we only use 256 neurons on the SYNTHETIC data set. For training Normalizing Flow in a supervised manner (OursSup) we used a learning rate of 5^{-7}, whilst the training in an end-to-end fashion (Ours) required a larger learning rate of 5^{-5} in order to perform well. All networks were trained for 400,000 iterations and validation error reported every 10,000 iterations. The model with the best validation error was chosen to compute test errors.

Outcome: Fig. 6 presents the qualitative results of the experiments on synthetic data. With the Inviwo software [30], we created two volume renderings of the same reconstruction, one with a cut through the middle of the volume and the second one with the complete volume. Furthermore, we show a single slice of the reconstructed volume. For the volume renderings we manually selected the transfer function on the ground truth volume to see all elements of the volume, and used it on the other methods.

We can see from the results that standard algorithms, WBP and SIRT, despite reconstructing the overall shape, miss fine details in certain areas. The methods are not able to cope with the noise and this results in a noisy 3D reconstruction. In L2Noisy, the noise is incorporated around the reconstructed shapes. However, the reconstruction is over-smoothed and small details are not well recovered. L2Den improves over L2Noisy and the noise around the volume disappears. Unfortunately, the fine details are still not recovered. When we look at both version of our method, OursSup and Ours, we see that the noise surrounding the volume disappears and the details are well reconstructed. We can see that both are close to the result obtained if there is no noise present on the observations, L2Clean. When we compare OursSup and Ours we see that Ours performs slightly better than the supervised version.

The main quantitative results are shown in Tab. 1. Here we see similar results as observed in the qualitative evaluation. WBP and SIRT obtain low performance in all metrics. L2Noisy performs better than standard algorithms but worse than L2Den. Ours achieves the best performance on all metrics. When compared with the supervised version of our method, OursSup, the latest achieves slightly better performance on the 2D metrics but worse than Ours in the 3D metrics. Lastly, the baseline L2Clean trained in the absence of noise, as expected, achieves the highest performance on almost all metrics. However, Ours obtains a better MSE on the 3D volume even if it is trained in the presence of noise. This indicates that our method is not only able to model the noise, but also that the reconstruction benefits from the unsupervised setup.

Lastly, we provide the results of the qualitative evaluation on the REAL data in Fig. 5. Here, we follow a similar procedure as in the SYNTHETIC data set and perform a volume-based visualization and a visual analysis on the slices of the reconstructed volume. We can see that the two baselines, WBP and SIRT, as in the synthetic data, incorporate the noise in the observation into the 3D volume, leading to reconstructions with low quality. On the other hand, Ours is able to recover high detailed volumes while L2Den generates an over-smoothed version as when applied to synthetic data. However, when compared to L2Noisy, even if Ours is able to better recover certain parts of the volume, the gap between these methods is smaller. Unfortunately, the lack of a ground truth volume makes it difficult to quantitatively determine which reconstructions is more accurate. Nevertheless, based on the qualitative evaluation and the results on synthetic data, we can conclude that Ours achieves a cleaner reconstruction.

Ablations: We evaluate how our framework performs under limiting model capacity for the reconstruction network. In this experiment, we reduce the number of features in the MLP from 256 to 32 and compare L2Noisy and Ours on our synthetic data set. We can observe that when the model capacity is reduced, both methods obtain a similar reconstruction. Ours achieved a MSE of 1.18 on the 3D volume while L2Noisy obtained 1.07. When the MSE is measure on the 2D images, we obtained 4.67 for Ours and 5.40 for L2Noisy. That might be an indication that the balance during training between 3D reconstruction and noise model requires a careful selection of the different hyper-parameters to separate 3D signal and noise.

Moreover, we evaluate the effect of accounting for the defocus in our image formation module. We use a synthetic dataset where we add a large defocus effect based on Eq. 5. We observed that accounting for this effect in the reconstruction process improves reconstruction accuracy, obtaining 0.91 MSE in the 3D volume instead 0.97 MSE when we do not account for it. For more detailed ablation studies we refer the reader to the supplementary material.

Limitations: Based on the ablation studies, the main limitation of our reconstruction algorithm is the careful selection of the hyper-parameters required to successfully separate 3D signal from noise. This problem might be tackled with computational resources as is done in other reconstruction software. However, we acknowledge that a large-scale evaluation of hyper-parameter selection on different data sets should be a future research direction.

5. Conclusion

We have shown, that a combination of a noise model and an implicit 3D shape representation can acquire 3D structure from noisy observations at a quality surpassing state of the art. To our knowledge, before no noise model for STEM was available and no implicit representation was fit to STEM imagery, in particular not jointly. Our combination makes progress along the most relevant access in this regime, the handling of noise and the representation of spatial detail. We would hope this approach will lend itself favourable to similar high-noise, non-photographic regimes with specific noise and image formation models.

Acknowledgments

This work was financed by the Baden-Württemberg Stiftung (BWS) for the ABEM project under grant MET-ID12–ABEM.

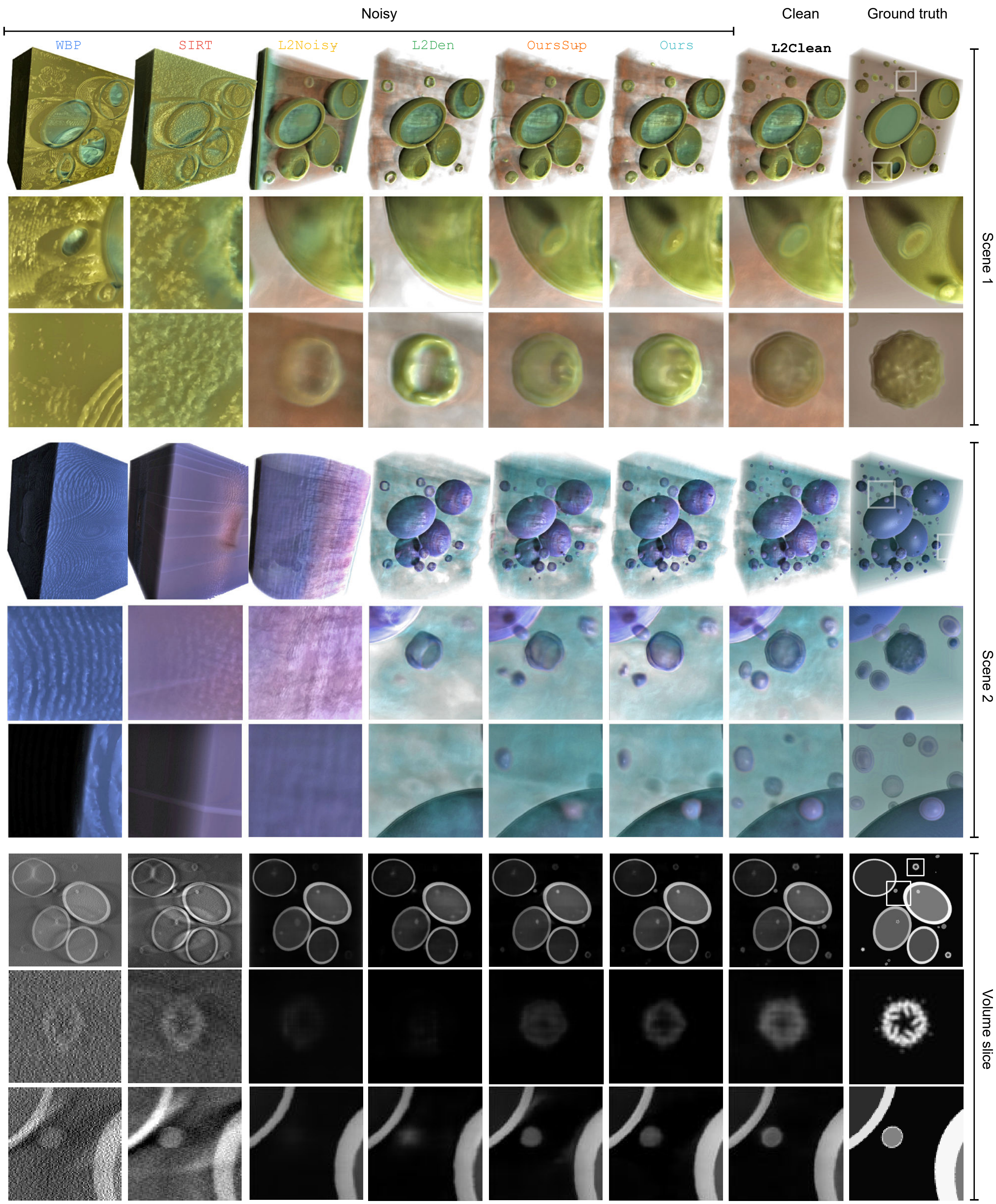


Figure 6. Qualitative results of different methods **(columns)** for different slices of the SYNTHETIC data set **(row triplets)**. In the presence of noise, WBP and SIRT cannot produce useful output. Training L2 on noisy images (L2Noisy) results in blurry details (in the top triplet) as well as strong constant bias over empty space. Training after denoising can remove this bias in empty space, but at the expense of spatial detail where the spherical virus structures have disappeared. Our OursSup and Ours resolve structures more similar to the ground truth.

References

[1] Deep learning STEM-EDX tomography of nanocrystals. *Nature Machine Intelligence*, 2021. 2

[2] Lodopab-ct, a benchmark dataset for low-dose computed tomography reconstruction. *Scientific Data*, 2021. 2

[3] Abdelrahman Abdelhamed, Marcus A Brubaker, and Michael S Brown. Noise flow: Noise modeling with conditional normalizing flows. In *ICCV*, pages 3165–73, 2019. 2, 4

[4] Michael D Abràmoff, Paulo J Magalhães, and Sunanda J Ram. Image processing with imagej. *Biophotonics international*, 11(7):36–42, 2004. 5

[5] V Abrishami, JR Bilbao-Castro, J Vargas, R Marabini, JM Carazo, and COS Sorzano. A fast iterative convolution weighting approach for gridding-based direct fourier three-dimensional reconstruction with correction for the contrast transfer function. *Ultramicroscopy*, 157:79–87, 2015. 2

[6] Jonas Adler and Ozan Öktem. Learned primal-dual reconstruction. *IEEE Trans. on Medical Imaging*, 2018. 2

[7] A.H. Andersen and A.C. Kak. Simultaneous algebraic reconstruction technique (sart): A superior implementation of the art algorithm. *Ultrasonic Imaging*, 1984. 2

[8] Rushil Anirudh, Hyojin Kim, Jayaraman J. Thiagarajan, K. Aditya Mohan, Kyle Champley, and Peer-Timo Bremer. Lose the views: Limited angle CT reconstruction via implicit sinogram completion. *CVPR*, 2017. 2

[9] Benjamin Attal, Eliot Laidlaw, Aaron Gokaslan, Changil Kim, Christian Richardt, James Tompkin, and Matthew O'Toole. Törf: Time-of-flight radiance fields for dynamic scene view synthesis, 2021. 2

[10] Daniel Otero Baguer, Johannes Leuschner, and Maximilian Schmidt. Computed tomography reconstruction using deep image prior and learned reconstruction methods. *Inverse Problems*, 2020. 2

[11] Jonathan T. Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. *ICCV*, 2021. 2

[12] Joshua Batson and Loic Royer. Noise2self: Blind denoising by self-supervision. In *ICML*, 2019. 2

[13] Tristan Bepler, Kotaro Kelley, Alex J. Noble, and Bonnie Berger. Topaz-denoise: general deep denoising models for cryoem and cryoet. *Nature Communications*, 2020. 2

[14] Adria Font Calvarons. Improved noise2noise denoising with limited data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021. 2

[15] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *CVPR*, pages 5939–48, 2019. 1, 4

[16] Albert V Crewe, J Wall, and J Langmore. Visibility of single atoms. *science*, 168(3937):1338–40, 1970. 1

[17] L. De Chiffre, S. Carmignato, J.-P. Kruth, R. Schmitt, and A. Weckenmann. Industrial applications of computed tomography. *CIRP Annals*, 2014. 2

[18] L. A. Feldkamp, L. C. Davis, and J. W. Kress. Practical cone-beam algorithm. *J OSA A*, 1984. 2

[19] Alessandro Foi, Mejdi Trimeche, Vladimir Katkovnik, and Karen Egiazarian. Practical poissonian-gaussian noise modeling and fitting for single-image raw-data. *IEEE Trans. on Image Processing*, 2008. 2

[20] Joachim Frank, editor. *Electron tomography: three-dimensional imaging with the transmission electron microscope*. 2013. 1, 2, 3

[21] Kyle Genova, Forrester Cole, Daniel Vlasic, Aaron Sarna, William T Freeman, and Thomas Funkhouser. Learning shape templates with structured implicit functions. In *ICCV*, pages 7154–7164, 2019. 2

[22] Kyle Genova, Forrester Cole, Avneesh Sud, Aaron Sarna, and Thomas A. Funkhouser. Local deep implicit functions for 3d shape. *CVPR*, 2020. 2

[23] Richard Gordon, Robert Bender, and Gabor T. Herman. Algebraic reconstruction techniques (art) for three-dimensional electron microscopy and x-ray photography. *J of Theoretical Biology*, 1970. 2

[24] Harshit Gupta, Thong H. Phan, Jaejun Yoo, and Michael Unser. Multi-cryogan: Reconstruction of continuous conformations in cryo-em using generative adversarial networks. In *ECCV*, 2020. 2

[25] Harshit Gupta, Michael T. McCann, Laurène Donati, and Michael Unser. Cryogan: A new reconstruction paradigm for single-particle cryo-em via deep adversarial learning. *IEEE Trans. on Computational Imaging*, 2021. 2

[26] Ji He, Yongbo Wang, and Jianhua Ma. Radon inversion via deep learning. *IEEE Trans. on Medical Imaging*, 2020. 2

[27] Peter Hedman, Pratul P. Srinivasan, Ben Mildenhall, Jonathan T. Barron, and Paul Debevec. Baking neural radiance fields for real-time view synthesis. *ICCV*, 2021. 2

[28] Richard Henderson. The potential and limitations of neutrons, electrons and x-rays for atomic resolution microscopy of unstained biological molecules. *Quarterly Rev of Biophysics*, 28(2):171–93, 1995. 1

[29] Philipp Henzler, Niloy J. Mitra, and Tobias Ritschel. Escaping plato's cave: 3d shape from adversarial rendering. In *ICCV*, October 2019. 1, 3

[30] Daniel Jönsson, Peter Steneteg, Erik Sundén, Rickard Englund, Sathish Kottravel, Martin Falk, Anders Ynnerman, Ingrid Hotz, and Timo Ropinski. Inviwo - a visualization system with usage abstraction levels. *IEEE Trans. on Visualization and Computer Graphics*, 2019. 7

[31] Earl J Kirkland. *Advanced computing in electron microscopy*, volume 12. 1998. 1

[32] Ivan Kobyzev, Simon Prince, and Marcus Brubaker. Normalizing flows: An introduction and review of current methods. *PAMI*, 2020. 1, 4

[33] James R. Kremer, David N. Mastronarde, and J. Richard McIntosh. Computer visualization of three-dimensional image data using imod. *J of Structural Biology*, 1996. 5

[34] Alexander Krull, Tim-Oliver Buchholz, and Florian Jug. Noise2void-learning denoising from single noisy images. In *CVPR*, 2019. 2

[35] Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. Noise2noise. *ICML*, 2018. 2

[36] Xinhao Liu, Masayuki Tanaka, and Masatoshi Okutomi. Practical signal-dependent noise parameter estimation from

a single noisy image. *IEEE Trans. on Image Processing*, 2014. 2

[37] Zhengchun Liu, Tekin Bicer, Rajkumar Kettimuthu, Doga Gursoy, Francesco De Carlo, and Ian Foster. Tomogan: low-dose synchrotron x-ray tomography with generative adversarial networks. *Journal of the Optical Society of America A*, 2020. 2

[38] Feng Long, Michael Doyle, Estefania Fernandez, Andrew S Miller, Thomas Klose, Madhumati Sevvana, Aubrey Bryan, Edgar Davidson, Benjamin J Doranz, Richard J Kuhn, et al. Structural basis of a potent human monoclonal antibody against zika virus targeting a quaternary epitope. *PNAS*, 116 (5):1591–1596, 2019. 5

[39] Roberto Marabini, Gabor T Herman, and José M Carazo. 3d reconstruction in electron microscopy using art with smooth spherically symmetric volume elements (blobs). *Ultramicroscopy*, 72(1-2):53–65, 1998. 2

[40] Nelson Max. Optical models for direct volume rendering. *IEEE Trans. on Visualization and Computer Graphics*, 1(2): 99–108, 1995. 3

[41] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *CVPR*, 2019. 2

[42] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *CVPR*, pages 4460–70, 2019. 1, 4

[43] Mateusz Michalkiewicz, Jhony K. Pontes, Dominic Jack, Mahsa Baktashmotlagh, and Anders Eriksson. Implicit surface representations as layers in neural networks. In *ICCV*, October 2019. 2

[44] 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*, pages 405–21, 2020. 1, 2, 3, 4

[45] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *CVPR*, 2019. 2

[46] Pawel Penczek, Michael Radermacher, and Joachim Frank. Three-dimensional reconstruction of single particles embedded in ice. *Ultramicroscopy*, 40(1):33–53, 1992. 2

[47] Stephen J Pennycook and Peter D Nellist. *Scanning transmission electron microscopy: imaging and analysis*. 2011. 1

[48] Michael Potmesil and Indranil Chakravarty. A lens and aperture camera model for synthetic image generation. *ACM SIGGRAPH Computer Graphics*, 15(3):297–305, 1981. 3

[49] Michael Radermacher. Weighted back-projection methods. In *Electron tomography*. 2007. 2

[50] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *ICML*, pages 1530–1538. PMLR, 2015. 1, 4

[51] Przemyslaw Rokita. Generating depth of-field effects in virtual reality applications. *IEEE Computer Graphics and Applications*, 16(2):18–21, 1996. 3

[52] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *ICCV*, pages 2304–14, 2019. 1

[53] C. O. S. Sorzano, J. Vargas, J. Otón, J. M. de la Rosa-Trevín, J. L. Vilas, M. Kazemi, R. Melero, L. del Caño, J. Cuenca, P. Conesa, J. Gómez-Blanco, R. Marabini, and J. M. Carazo. A survey of the use of iterative reconstruction algorithms in electron microscopy. *BioMed Research International*, 2017. 2

[54] Richard Szeliski. *Computer vision: algorithms and applications*. 2010. 1

[55] Towaki Takikawa, Joey Litalien, Kangxue Yin, Karsten Kreis, Charles Loop, Derek Nowrouzezahrai, Alec Jacobson, Morgan McGuire, and Sanja Fidler. Neural geometric level of detail: Real-time rendering with implicit 3D shapes. 2021. 2

[56] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Carsten Stoll, and Christian Theobalt. Patchnets: Patch-based generalizable deep implicit 3d shape representations. In *ECCV*, 2020. 2

[57] Shubham Tulsiani, Tinghui Zhou, Alexei A Efros, and Jitendra Malik. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *CVPR*, pages 2626–2634, 2017. 1

[58] Barnali Waugh, Sharon G. Wolf, Deborah Fass, Eric Branlund, Zvi Kam, John W. Sedat, and Michael Elbaum. Three-dimensional deconvolution processing for STEM cryotomography. *Proceedings of the National Academy of Sciences*, 2020. 2

[59] Qingsong Yang, Pingkun Yan, Yanbo Zhang, Hengyong Yu, Yongyi Shi, Xuanqin Mou, Mannudeep K Kalra, Yi Zhang, Ling Sun, and Ge Wang. Low-dose ct image denoising using a generative adversarial network with wasserstein distance and perceptual loss. *IEEE Trans. on medical imaging*, 2018. 2

[60] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *CVPR*, 2021. 2

[61] Guangming Zang, Ramzi Idoughi, Congli Wang, Anthony Bennett, Jianguo Du, Scott Skeen, William L. Roberts, Peter Wonka, and Wolfgang Heidrich. Tomofluid: Reconstructing dynamic fluid from sparse view videos. In *CVPR*, June 2020. 2

[62] Guangming Zang, Ramzi Idoughi, Rui Li, Peter Wonka, and Wolfgang Heidrich. Intratomo: Self-supervised learning-based tomography via sinogram synthesis and prediction. *ICCV*, 2021. 2

[63] Yide Zhang, Yinhao Zhu, Evan Nichols, Qingfei Wang, Siyuan Zhang, Cody Smith, and Scott Howard. A poisson-gaussian denoising dataset with real fluorescence microscopy images. In *CVPR*, 2019. 2

[64] Ellen D. Zhong, Tristan Bepler, Bonnie Berger, and Joseph H. Davis. Cryodrgn: reconstruction of heterogeneous cryo-em structures using neural networks. *Nature Methods*, 2021. 2

[65] Ellen D. Zhong, Adam Lerer, Joseph H. Davis, and Bonnie Berger. Cryodrgn2: Ab initio neural reconstruction of 3d protein structures from real cryo-em images. In *ICCV*, 2021.

2
[66] Bo Zhu, Jeremiah Z. Liu, Stephen F. Cauley, Bruce R. Rosen, and Matthew S. Rosen. Image reconstruction by domain-transform manifold learning. *Nature*, 2018. 2