Self-Supervised Cryo-Electron Tomography Volumetric Image Restoration from Single Noisy Volume with Sparsity Constraint

Zhidong Yang[1,2,3], Fa Zhang[1,*], Renmin Han[2,*]
[1]High Performance Computer Research Center, ICT, CAS
[2]Research Center for Mathematics and Interdisciplinary Sciences, Shandong University
[3]University of Chinese Academy of Sciences

Abstract

Cryo-Electron Tomography (cryo-ET) is a powerful tool for 3D cellular visualization. Due to instrumental limitations, cryo-ET images and their volumetric reconstruction suffer from extremely low signal-to-noise ratio. In this paper, we propose a novel end-to-end self-supervised learning model, the Sparsity Constrained Network (SC-Net), to restore volumetric image from single noisy data in cryo-ET. The proposed method only requires a single noisy data as training input and no ground-truth is needed in the whole training procedure. A new target function is proposed to preserve both local smoothness and detailed structure. Additionally, a novel procedure for the simulation of electron tomographic photographing is designed to help the evaluation of methods. Experiments are done on three simulated data and four real-world data. The results show that our method could produce a strong enhancement for a single very noisy cryo-ET volumetric data, which is much better than the state-of-the-art Noise2Void, and with a competitive performance comparing with Noise2Noise. Code is available at https://github.com/icthrm/SC-Net.

1. Introduction

Cryo-ET is a powerful technique for the visualization of cellular ultrastructure and macromolecules in three-dimensional space, where 3D structures can be reconstructed from a series of 2D images (tilt series) taken from different angles [6, 9]. However, due to instrumental limitations, cryo-ET volumetric images always suffer from extremely low Signal-to-Noise Ratio (SNR). To restore the corrupted ultrastructure from noisy 3D volume is an essential task in cryo-ET data analysis.

Recently, trainable Deep Neural Network (DNN) based image restoration model has attracted much attention because of its excellent performance, which can be divided into two categories: clean-target-based supervised learning model and self-supervised learning model. The supervised learning model needs plenty of "noisy-clean" image pairs to train a reliable model. Although supervised learning model is able to achieve very good performance with clearly defined datasets, it cannot well handle the real-world noisy data whose ground-truth or noisy pattern is unavailable, for example, the cryo-ET data. To improve this, self-supervised model is proposed. Self-supervised model does not need ground-truth information during training, in which the supervisory information is observed from the original noisy data. Recently, Noise2Noise (N2N) [19] and Noise2Void (N2V) [17] have emerged as two main branches of self-supervised restoration model. Specifically, N2V is able to perform single-image training. However, for the images with very high noise, the results produced by N2V may still be too noisy, especially for the 3D volumetric image.

Here, we mainly focus on the image restoration of volumetric data in cryo-Electron Tomography (cryo-ET). One possible data enhancement strategy is to first filter the 2D projections and then build a denoised 3D tomographic volume [21]. Although such procedure can produce a smoothed tomography, its output is usually over-smoothed and loses a lot of fine-grained structures. Because the filtering operation on 2D projections cannot strictly keep the 2D-to-3D relationship defined by Fourier-slice-theorem [24], it is more reasonable to directly operate on the 3D cryo-ET volumetric image [6]. However, it is very hard to acquire a large number of homogeneous volumetric images for a specific kind of specimen in cryo-ET, which hampers the application of restoration models relied on large training dataset.

In this paper, we propose a self-supervised deep learning model, the Sparsity Constrained Network (SC-Net), to directly restore cryo-ET 3D volumetric image from only a single noisy data. By combining the structure information from raw noisy volume and local smoothness information from smoothed representation, our proposed model can suppress high-frequency noise and preserve ultrastructure details in a self-supervised manner, requiring only one single

*All correspondence should be addressed to Fa Zhang (zhangfa@ict.ac.cn) and Renmin Han (hanrenmin@sdu.edu.cn).

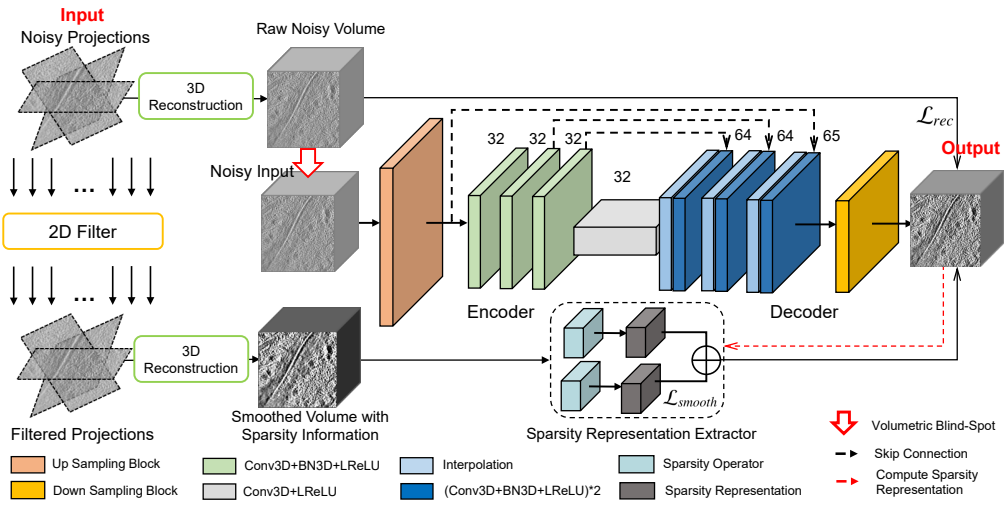


Figure 1: The Architecture of SC-Net. The workflow starts with a set of projections, which will be (i) reconstructed to generate a raw noisy volumetric image, (ii) filtered and reconstructed to generate a smoothed volumetric image, respectively. SC-Net adapts a blind-spot replacement on the noisy volume for model training and utilizes the local smooth (sparsity) information extracted from the smoothed volume to guide the training.

noisy data as input during training. Experiments done on both simulated and real-world datasets show that SC-Net can produce strong enhancement for noisy input, much better than the ones produced by Noise2Void trained on single noisy input and competitive with the ones produced by Noise2Noise (3D) trained on hundreds of tomograms.

The main contributions are as follows:

- A self-supervised volumetric image restoration architecture is proposed, by adapting a blind-spot approach with sparsity constraint extracted from a self-lifted representation of the input data (shown in Figure 1).
- Up-sampling block is first introduced to self-supervised restoration network as a module for augmentation to preserve structure details and prevent the output from being over-smoothed.
- A new combined loss function is devised in our image restoration model, in which the sparsity information in self-lifted representation is used to enforce local smoothness in structure restoration.
- A general procedure for the simulation of electron tomographic photographing and imaging is proposed, for the ease of quantitative analysis.

2. Related Work

The restoration of degraded image is an essential task for the data analysis in cryo-ET, which is also one of the main problems in the field of computer vision.

2.1. Traditional Volumetric Image Restoration

Several traditional filtering methods have been adapted for the restoration of volumetric image in cryo-ET, such as the multiscale transformation [29], non-linear anisotropic diffusion [8, 7], bilateral filtering [14, 25], and iterative median filtering [31]. These methods work on the reconstructed tomographic data, trying to suppress the noise and strengthen the ultrastructure with predefined smooth assumption. On the contrary, reconstruction methods with specified filtering are proposed to improve volumetric image quality by introducing the predefined smooth or sparsity assumption in the reconstruction procedure [5, 33]. Non-local methods such as BM4D [3, 4, 20] have also been applied in 3D volumetric image denoising and restoration and show quite good performance. The non-local methods introduce the idea of "Grouping" by block-matching and collaborative filtering, assuming that the expectancy of the noise will converge to zero when enough image patches are collected and reweighted.

2.2. Learning-based Image Restoration

At the beginning, learning-based image restoration methods are proposed in a supervised fashion. Early in 2008, neural network has already been applied to image denoising and restoration [13]. Then, with the development of deep neural network (DNN), a neural network model called DnCNN is proposed [34], which is based on residual blocks [12] and trained on a set of "noisy-clean" image pairs, showing excellent performance in image restoration. In the same year, the encoder-decoder network is introduced to image denoising and shows its effectiveness [22]. Following these works, noise prior is introduced in supervised training to enhance model performance on real-world noisy data [35, 10]. In recent times, inspired by non-local network [32], non-local blocks are adapted to image restoration model, achieving a good performance on image denoising, super-resolution, and other related tasks [36].

Although supervised methods have achieved great success, its model training requires clean image as reference, which is hard to be satisfied when noise model is unknown. To overcome this problem, several self-supervised approaches have been proposed. [30] firstly states that a generator network is sufficient to capture plenty of low-level image statistics prior for any learning. By applying statistical reasoning to signal reconstruction, Noise2Noise realizes image restoration with only corrupted examples [19]. Given the known noise distribution and intensity, an enhanced N2N approach is proposed [23]. To capture the statistic information, a number of training data are required in N2N. Meanwhile, an image training procedure with blind-spot replacement is proposed in Noise2Void [17], leading to a diversity of self-supervised models [1, 18, 27]. Especially, the blind-spot replacement approach supports model training with only a single noisy image.

Inspired by Noise2Noise [19], learning-based image restoration has also been introduced to cryo-ET, i.e., the Topaz-Denoiser [2]. Similar to other N2N methods, the training of Topaz also requires a large dataset. Nevertheless, unexpected artifacts usually occur for a specific image which is quite different from images in training dataset.

3. Preliminaries

To simplify the discussion, we assume that a projection image $\boldsymbol{I}_n(x, y)$ in cryo-ET is a discrete observation of the projection $\boldsymbol{P}_n(x, y)$ with additive Gaussian noise $\boldsymbol{N}(x, y)$, i.e. $\boldsymbol{I}_n = \boldsymbol{P}_n + \boldsymbol{N}_n$. Consequently, we have:

Lemma. The additive Gaussian noise in 2D projection remains as Gaussian noise in 3D reconstruction.

$$\boldsymbol{V}(\boldsymbol{x}) = \boldsymbol{\Phi}(\boldsymbol{x}) + \boldsymbol{N}(\boldsymbol{x}), \tag{1}$$

where $\boldsymbol{V}(\boldsymbol{x})$ is notated as a volumetric image reconstructed from series of $\boldsymbol{I}_n$, $\boldsymbol{\Phi}(\boldsymbol{x})$ is the ideally clean image reconstructed from $\boldsymbol{P}_n$ and $\boldsymbol{N}(\boldsymbol{x})$ is noise in 3D space. This is the theoretical basis of our SC-Net. Detailed proof can be found in Supplementary Material S1.

4. Method

According to Section 3, a pseudo clean volume can be obtained by suppressing the noise in projections before reconstruction, providing an estimation of the smoothness and sparsity of the ideal 3D volumetric image, which is the basic idea of our Sparsity Constrained Network (SC-Net).

4.1. Process Overview

The workflow starts with the projections of a cryo-ET data, which will be (i) reconstructed to generate a raw noisy volumetric image, (ii) filtered and reconstructed to generate a smoothed volumetric image, respectively. Then, these two volumes will be fed into SC-Net.

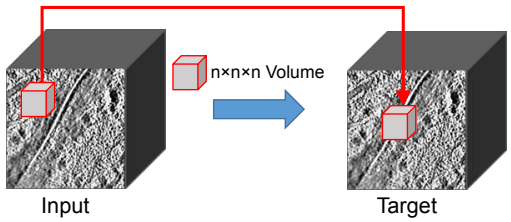


Figure 2: Volumetric blind-spot replacement. The $n \times n \times n$ region around an interested voxel is taken into account in blind-spot training instead of one single voxel.

4.2. Network Architecture

SC-Net is a UNet-based encoder-decoder network [28], whose detailed architecture is shown in Figure 1. SC-Net adapts a 3-depth encoder and decoder network. An up-Sampling Block ($2\times$ up-sampling) before the encoder layers is adapted to lift the receptive field and protect structure information from over-smooth. A basic convolutional block consists of a 3^3 Conv3D layer with a stride of 2, a Batchnorm3D layer and a LReLU with $k = 0.1$. The up-sampling block adapts a $2\times$ nearest interpolation and the down-sampling block adapts a $0.5\times$ nearest down-sampling.

SC-Net accepts the raw noisy volumetric image and the smoothed volumetric image as input. These two volumes will then be clipped into a set of corresponding small overlapped patches (L^3 voxels for each patch) respectively. A volumetric blind-spot replacement strategy is applied on the noisy volume patches for model training, and a sparsity representation extractor is specifically designed to extract the local smooth information from the smoothed volume patches to guide the training.

4.3. Main Components

Smoothed Volume Generation. A 2D filter is adapted on the input projections to produce a smoothed volumetric image $\hat{\boldsymbol{V}}(\boldsymbol{x})$ which can provide sparsity information to represent smoothness in an ideal volumetric image.

Volumetric Blind-spot Replacement. Cryo-ET volumetric image is composed of strong coherent ultrastructures, which makes single-pixel blind-spot strategy proposed in Noise2Void unsuitable. Here, a volumetric blind-spot strategy is devised. At each training iteration, the volumetric blind-spot approach randomly selects a set of voxels from the volumetric patch $\boldsymbol{V}(\boldsymbol{x})$. For each selected voxel, its $(n-1)$-neighbouring voxels will be taken into account in the receptive field together with the selected voxel, replaced by a $n \times n \times n$ region randomly selected from its local neighbouring area (as shown in Figure 2). The replaced volumetric patch $\boldsymbol{V}^{bs}(\boldsymbol{x})$ along with the raw input patch $\boldsymbol{V}(\boldsymbol{x})$ compose a training pair at each iteration.

Sparsity Representation Extractor. The sparsity information residing in the smoothed volume could provide a guidance in image restoration. A gradient (first-order

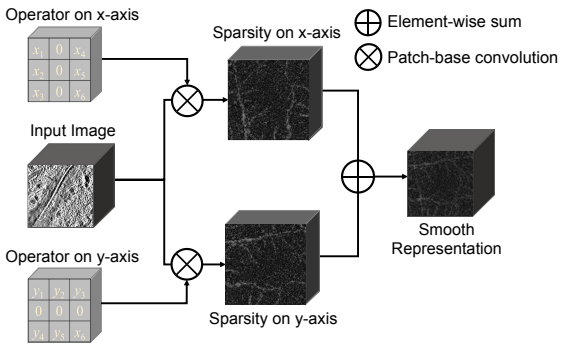


Figure 3: Details of sparsity representation extractor. Patch convolution with Sobel kernel is used to extract the sparsity.

derivative) based sparsity representation extractor is designed to capture these supervision information (as shown in Figure 3). The sparse representation extracted by Sobel kernel will be integrated and fed into the network to guide the training.

4.4. Loss Function

A combined loss function is introduced to SC-Net for both noise suppression and structure preservation.

Volumetric Reconstruction Loss. Due to the lack of training data, it is hard for a network to converge to clean signal in a low-shot self-supervised training. Based on the blind-spot strategy, a volumetric reconstruction loss function is introduced here, i.e.,

$$\mathcal{L}_{rec} = ||\boldsymbol{M} \odot f_\theta(\boldsymbol{V}_i^{bs}(\boldsymbol{x})) - \boldsymbol{M} \odot \boldsymbol{V}_i(\boldsymbol{x})||_2^2, \tag{2}$$

where $\boldsymbol{V}_i^{bs}(\boldsymbol{x})$ is the patch obtained by blind-spot replacement on the i^{th} image patches, $f_\theta(\boldsymbol{V}_i^{bs}(\boldsymbol{x}))$ is the output of our network model, $\odot$ is the Hadamard product and $\boldsymbol{M}$ is a 3D mask to indicate whether a voxel (x, y, z) in $\boldsymbol{V}_i^{bs}(\boldsymbol{x})$ has a blind-spot replacement or not,

$$\boldsymbol{M}_{(x,y,z)} = \begin{cases} 1 & \text{if } \boldsymbol{V}_i^{bs}(x,y,z) \text{ has been replaced,} \\ 0 & \text{if } \boldsymbol{V}_i^{bs}(x,y,z) \text{ has not been replaced.} \end{cases} \tag{3}$$

That is, here we only focus on the replaced voxels after blind-spot replacement, which makes our model preserve the structure information from the noisy input but not converge to the noise.

Sparsity-Guided Smoothing Loss. The first-order derivative of an image is sensitive to extreme changes. Hence it will produce a sparse representation if an image is smooth enough. Based on this truth, we introduce a first-order derivative-based loss function named sparsity-guided smoothing loss to transfer the sparsity of smoothed image into output image.

We adapt Sobel operator as a function to map network output and smoothed volume $\hat{\boldsymbol{V}}(\boldsymbol{x})$ to their first-order derivative separately, which is formulated as

$$\mathcal{D}_s(\hat{\boldsymbol{V}}(\boldsymbol{x})) = |G_x \otimes \hat{\boldsymbol{V}}(\boldsymbol{x})| + |G_y \otimes \hat{\boldsymbol{V}}(\boldsymbol{x})|, \tag{4}$$

where $\otimes$ is notated as convolution operator, G_x and G_y are the Sobel operator on x-axis and y-axis separately. Thus, the sparsity-guided smoothing loss can be defined as

$$\mathcal{L}_{smooth} = ||\mathcal{D}_s(f_\theta(\boldsymbol{V}_i^{bs}(\boldsymbol{x}))) - \mathcal{D}_s(\hat{\boldsymbol{V}}(\boldsymbol{x}))||_2^2. \tag{5}$$

Expectancy Constraint Loss. To overcome the mean shifting problem (an unstable shifting in pixel value distribution) in blind-spot inference trained with insufficient data, a loss function named expectancy constraint loss is devised to let the expectancy of the output image approximate to the one of smoothed image.

Firstly, we need to compute mean of network output and that of smoothed image by patches with kernel size of S^3 which can represents local smoothness. We name this as *local expectancy*, notated as $\mu_{\boldsymbol{V}}$. This is formulated as

$$\mu_{\boldsymbol{V}} = \boldsymbol{V}(\boldsymbol{x}) \otimes \boldsymbol{1}_{\boldsymbol{S}}, \tag{6}$$

where $\otimes$ is convolution operator, $\boldsymbol{1}_{\boldsymbol{S}}$ is a convolutional kernel with size of S^3 and all elements are $\frac{1}{S^3}$.

Analogously, we calculate the *global expectancy* of μ_{f_θ} and $\mu_{\hat{\boldsymbol{V}}}$, both of which are computed with all available pixels in an image, and notate these two global expectancy as $\mathbb{E}(\mu_{f_\theta})$ and $\mathbb{E}(\mu_{\hat{\boldsymbol{V}}})$, respectively. Finally, $\mathbb{E}(\mu_{f_\theta})$ and $\mathbb{E}(\mu_{\hat{\boldsymbol{V}}})$ will be the input of expectancy constraint loss function. Detailed formulation can be defined as

$$\mathcal{L}_{exp} = ||\mathbb{E}(\mu_{f_\theta}) - \mathbb{E}(\mu_{\hat{\boldsymbol{V}}})||_2^2. \tag{7}$$

Combined Loss Function. Additionally, a regularization loss $\mathcal{L}_{reg} = ||\nabla f_\theta(\boldsymbol{V}_i^{bs}(\boldsymbol{x}))||_1$ (∇ is the first-order gradient of an image) is introduced in our loss function to prevent over-fitting during training. And the complete loss function of our model is define as

$$\mathcal{L} = \lambda_1 \mathcal{L}_{rec} + \lambda_2 \mathcal{L}_{smooth} + \lambda_3 \mathcal{L}_{exp} + \lambda_4 \mathcal{L}_{reg}. \tag{8}$$

5. Experiments

We evaluated the performance of SC-Net on three simulated datasets and four real Cryo-ET datasets, and compared our method with Topaz-Denoiser 3D (a 3D version of Noise2Noise) [2], 3D version of Noise2Void [19], BM4D [20] and low-pass filter, resulting in a detailed quantitative empirical analysis. The sources of datasets can be found in Supplementary Material S2.

5.1. Network Training Details

SC-Net was implemented by PyTorch [26]. For all the experiments, the model was trained on two NVIDIA GTX 2080 GPUs, online for single noisy data each time. The batch size was set to 2 with 128^3 patch size during training. The model was trained by 15 epochs for each noisy data, where the optimizer is Adam [15] with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The learning rate was set to 0.0004. For Eq.

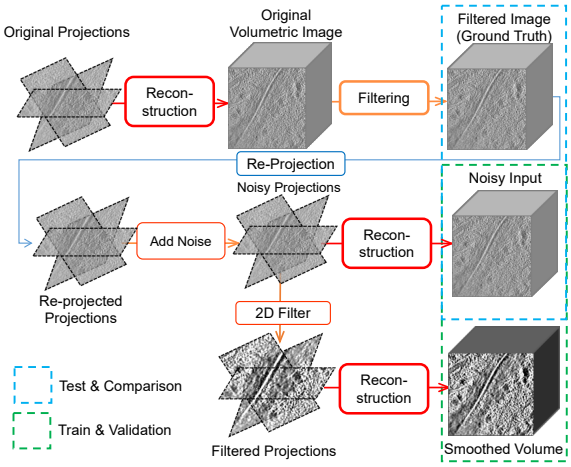


Figure 4: The workflow to prepare a simulated dataset. The 2D filter keeps the same with the one in SC-Net.

8, the parameters were set as $\lambda_1 = 0.6$, $\lambda_2 = 0.1$, $\lambda_3 = 0.2$, and $\lambda_4 = 0.01$ when training with L_{exp}, and set as $\lambda_1 = 0.8$, $\lambda_2 = 0.1$, $\lambda_3 = 0.0$, and $\lambda_4 = 0.01$ when training without L_{exp}. For 3D reconstruction, we used the *tilt* program in IMOD [16] to get volumetric image from 2D projections. We used a pretrained model of Topaz Denoiser [2] trained with a large-scale image volume dataset as our 2D filter. During model training, the size of volumetric blind-spot replacement was set to $3 \times 3 \times 3$, and the ratio of replaced voxels in a volumetric patch was set to 10%.

5.2. Preparation of Simulated Dataset

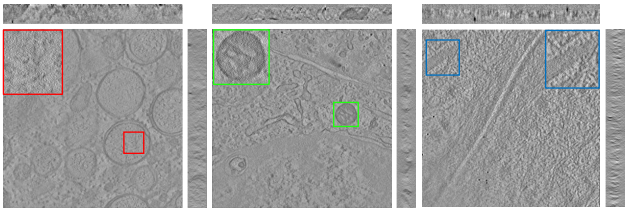
A novel procedure for the preparation of simulated dataset with ground-truth is proposed (shown in Figure 4). Given a raw tilt series, a noisy volumetric image can be reconstructed from this tilt series. Here, the *tilt* program in IMOD [16] is used to get the volumetric image (other similar software [11] is also selectable for the volumetric reconstruction). Then a Gaussian filter (notated as "Filtering" in Figure 4) with $\sigma = 2$ is applied to this noisy volumetric image. The filtered image is regarded as *ground-truth volume*.

After this, we re-project this ground-truth volume to get a series of reprojections which will then be normalized and regarded as the *ground-truth projections*. Thus, different kinds of additive noise can be added to ground-truth projections to provide simulation with different noise levels. Here, it should be noted that the noise is added to the 2D projections but not reconstructed 3D volume.

5.3. Experiments on Simulated Data

The experiments on simulated dataset were conducted on three data[1], named SARS-ConoraVirus, Synapse and Adhesion Belt. We mainly evaluated the image restoration performance on additive white Gaussian noise (AWGN) with intensity $\sigma = 10$, 15 and 20. Figure 5 shows the noisy reconstruction examples with AWGN ($\sigma = 20$), where the demonstrated images are selected from the middle slice of the tomograms along the direction of x-, y- and z-axis. Details about the visualization rules can be found in Supplementary Material S4.

SC-Net were compared with the other single-image-based methods and Topaz (trained on large dataset). Figure 6 shows the visual results of these methods[2], while Table 1 and Table 2 show the quantitative analysis on these results. Judging from Figure 6, we can find that the results of SC-Net are very close to the ground-truth. Judging from the PSNR values in Table 1, we can find that the BM4D and SC-Net perform the best among these five methods, and BM4D performs the best in most cases. It is reasonable that the



SARS 18.27/0.189 Synapse 21.34/0.238 Belt 18.64/0.212

Figure 5: Examples of the noisy reconstructed volumetric data (metrics: PSNR (dB)/SSIM), in which the volumes are reconstructed from projections with AWGN ($\sigma = 20$).

Table 1: PSNR results on simulated dataset. For each column, the **top** and **second** values are highlighted.

Methods	Belt			Synapse			SARS		
	10	15	20	10	15	20	10	15	20
Noisy	22.76	23.44	18.64	27.79	26.02	21.34	23.72	22.75	18.27
Topaz	28.06	27.90	25.63	31.72	32.85	27.53	30.1	33.07	29.39
BM4D	**29.36**	**30.11**	**28.61**	**31.06**	**36.02**	**32.90**	**36.32**	**34.46**	**32.36**
LPF	25.84	22.39	16.93	29.05	27.24	24.55	24.18	24.14	17.97
N2V	23.38	25.63	21.87	26.48	17.97	21.29	22.48	22.26	17.35
Ours (no L_{exp})	**32.38**	**27.36**	**26.46**	**32.33**	**28.58**	**31.73**	32.23	**31.01**	27.58
Ours	28.61	25.68	23.41	30.03	27.55	29.02	**34.64**	27.36	**32.79**

Table 2: SSIM results on simulated dataset. For each column, the **top** and **second** values are highlighted.

Methods	Belt			Synapse			SARS		
	10	15	20	10	15	20	10	15	20
Noisy	0.423	0.282	0.212	0.503	0.316	0.238	0.401	0.270	0.189
Topaz	0.979	0.974	0.968	0.990	0.986	0.980	0.970	0.960	0.950
BM4D	0.798	0.662	0.541	**0.950**	0.897	0.829	0.883	0.793	0.682
LPF	0.470	0.329	0.266	0.575	0.388	0.269	0.481	0.345	0.234
N2V	0.547	0.389	0.304	0.475	0.355	0.239	0.385	0.261	0.179
Ours (no L_{exp})	**0.922**	**0.910**	**0.899**	0.928	**0.904**	**0.840**	**0.905**	**0.886**	**0.780**
Ours	**0.914**	**0.899**	**0.870**	**0.932**	**0.916**	**0.860**	**0.906**	**0.886**	**0.852**

[1]All the quantitative analysis on the simulated volumetric image are measured with PSNR (dB)/SSIM (Supplementary Material S3).

[2]Enlarged results are available in Supplementary Material S5.1.

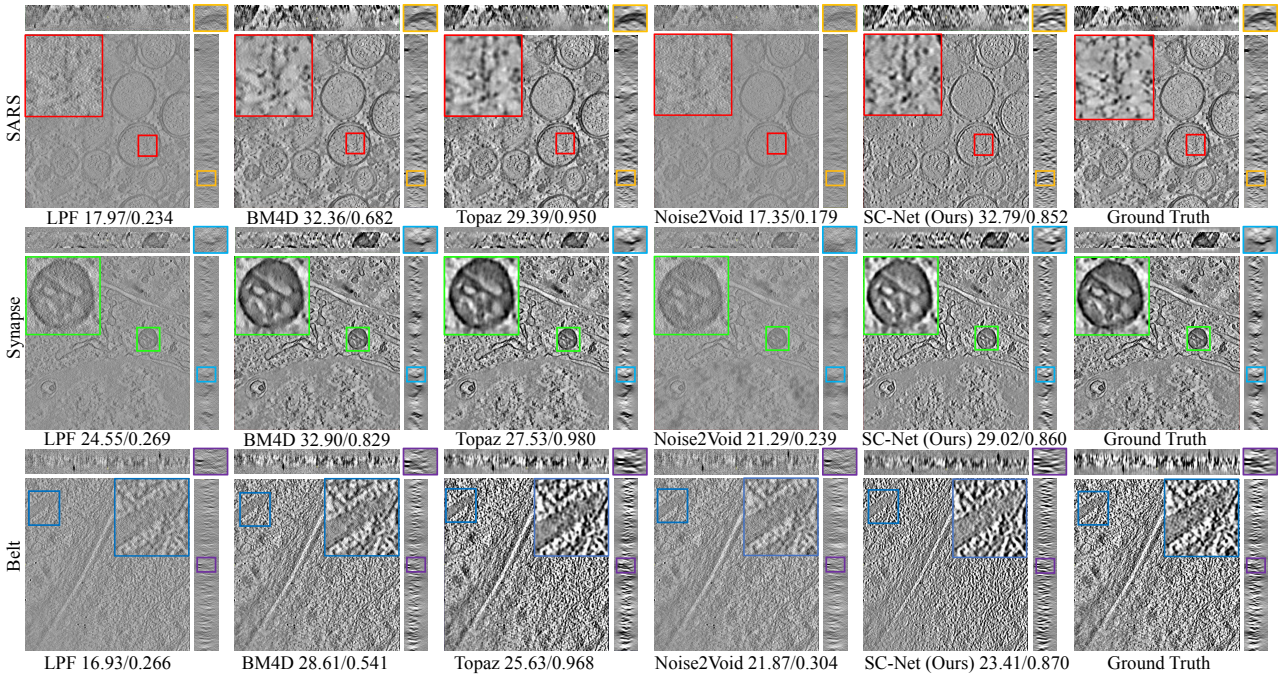


Figure 6: Results of the simulated data with AWGN ($\sigma = 20$) shown in 3D space (metrics: PSNR(dB)/SSIM).

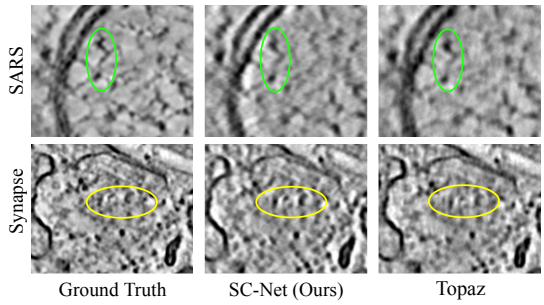


Figure 7: Local enlarged results of the simulated data with AWGN ($\sigma = 20$).

simulation is done with AWGN and the Wiener filter used in BM4D is specially optimized for AWGN. Judging from the SSIM values in Table 2, we can find that scores of SC-Net rank the best among single-image-based methods. In addition, the introduction of expectancy constraint makes a contribution to structure enhancement, as the improvement of SSIM values shown in Table 2.

It is promising that SC-Net outperforms the other single-image-based N2V methods in both quantitative analysis and visual comparison, which means that the introduction of sparsity-guided smoothing loss can improve restoration performance of blind-spot inference. Also, SC-Net can provide more stable structure preservation than BM4D.

5.4. Comparisons between Topaz

Topaz is a 3D N2N restoration model trained on hundreds of tomograms. Quantitative results in Table 2 show that Topaz performs better on SSIM value because Topaz is a large data-driven model, with which enough smoothing representation is possible to be learned. However, SC-Net still performs better than Topaz on PSNR in most cases, which means that our SC-Net has strong ability in structure restoration even trained with single noisy image. For data-driven methods, unexpected artifacts usually occur for specific images which are different from images in training dataset. Figure 7 gives such examples, which shows obvious grid artifacts in the results from Topaz. Additionally, training cost of our SC-Net is significantly lower than Topaz as we implement single-image training.

5.5. Real-world Datasets

Four real-world datasets were tested in our experiment: Centriole, Mitochondria, Vesicle and VEEV. **Centriole** is a tilt series of 64 projections ranging from -61.0° to +65.0° at 2° intervals. The size of each tilt image is 1024×1024 px with 1.01 nm/px. **Mitochondria** is a tilt series of 120 projections ranging from $-52.0°$ to $59.0°$ at $1°$ interval. The size of each tilt image is 1024×1024 px with 0.8 nm/px. **Vesicle** is a tilt series of 120 projections ranging from $-59.0°$ to $+60.0°$ at a $1°$ interval. The size of each tilt image is 1024×1024 px with 0.8 nm/px. **VEEV** is a tilt series of 21 projections ranging from -50.0° to +50.0° at 5° intervals. The size of each tilt image is 1536×2048 px with 0.2 nm/px. The tomographic volumes for these specimens were reconstructed by the *tilt* program in IMOD [16].

5.6. Experiments on Real Cryo-ET Data

Figure 8 shows the visual results of the experiments on real cryo-ET datasets, where the 3D-style visual results are

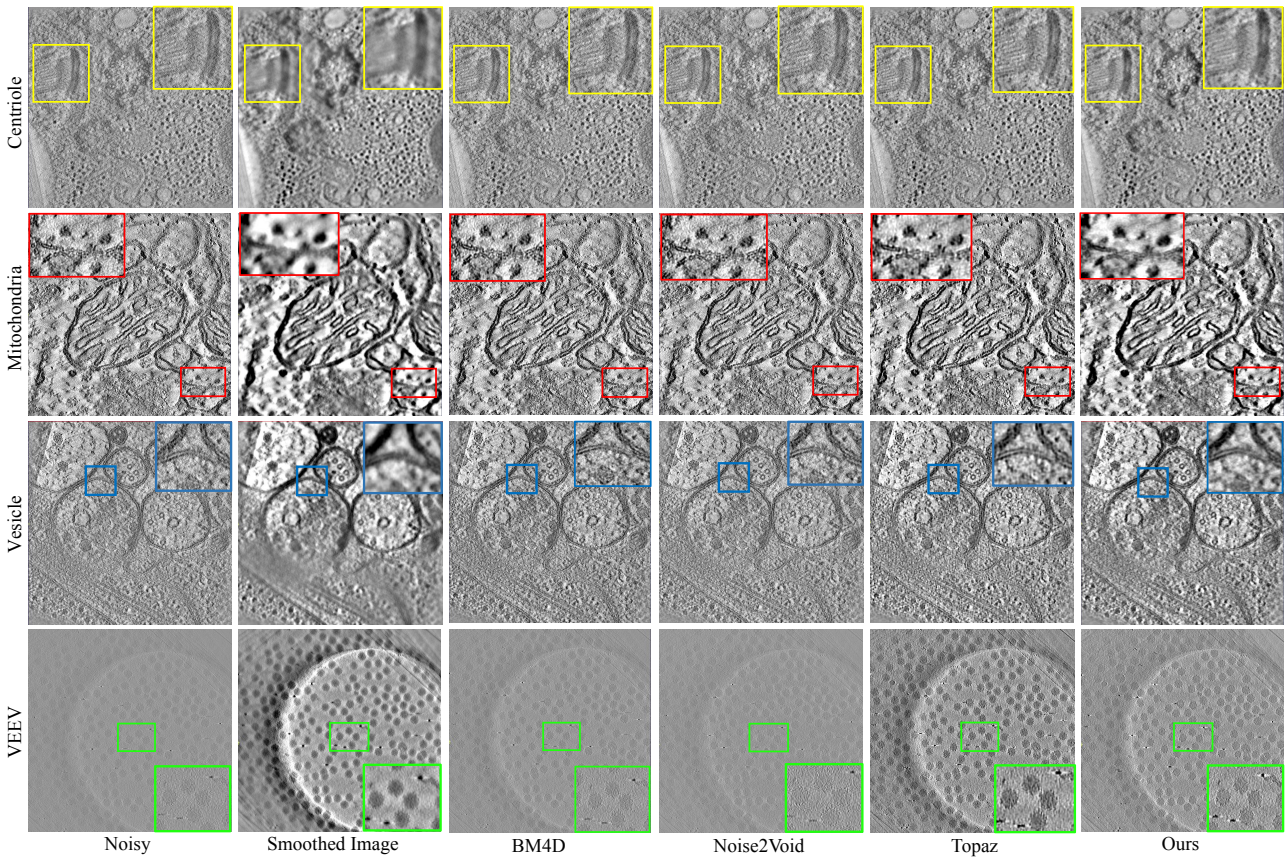


Figure 8: Visual results of the real data on y-slice.

provided in Supplementary Material S5.2. Judging from the visual results, we can find that SC-Net performs almost the best. And both SC-Net and Topaz (N2N) perform better than BM4D. It is reasonable for the performance degeneration of BM4D on real cryo-ET data, in which the Wiener filter is not robust to real-world complex noise.

As ground-truth is unavailable in real-world situations, we decide to evaluate the performance of SC-Net by $\text{FSC}_{e/o}$. The Fourier shell correlation comparison between two volumetric images calculated from the even and odd projection images respectively ($\text{FSC}_{e/o}$) is a popular resolution measure in the field of cryo-ET. $\text{FSC}_{e/o}$ is adapted from FSC. The definition of FSC is as follows:

$$\text{FSC}(r) = \frac{\sum_{r_i \in r} F_1(r_i) \cdot F_2(r_i)^*}{\sqrt{\sum_{r_i \in r} |F_1(r_i)|^2 \sum_{r_i \in r} |F_2(r_i)|^2}}, \tag{9}$$

where F_1 is complex factor for volume 1, F_2^* conjugate of the structure Factor for volume 2, and r_i is the individual voxel element at radius r.

Assuming that SNR for each map from a half reconstruction is with half signals of that of the full reconstruction, $\text{FSC}_{e/o}$ is calculated as

$$\text{FSC}_{e/o}(r) = \frac{2\text{FSC}(r)}{\text{FSC}(r) + 1}. \tag{10}$$

Table 3: Resolution estimated by $\text{FSC}_{e/o}^{-1}(0.5)$ (Angstrom, Å). For resolution of a tomogram, the lower is the better.

Dataset	Vesicle	Mitochondria	Centriole	VEEV
Noisy	30.06	23.05	57.74	9.36
BM4D	29.44	22.32	57.37	9.35
Topaz	30.42	23.09	55.16	**6.8**
N2V	30	23.02	57.67	9.36
Ours	**22.66**	**14.80**	**38.21**	8.31

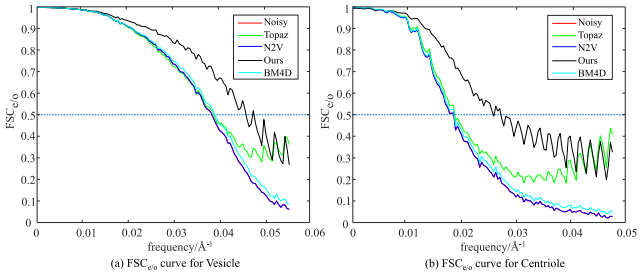


Figure 9: Examples of $\text{FSC}_{e/o}$ curves. Blue dash line in the figure points out the position of $r = 0.5$.

Table 3 shows the $\text{FSC}_{e/o}^{-1}(0.5)$ resolution calculated on the four real-world data. Except for the $\text{FSC}_{e/o}^{-1}(0.5)$ of VEEV, which is calculated with subtomogram sized by $60^2 \times 100$, all the others are calculated with subtomograms sized by $512^2 \times 100$. Results show that SC-Net achieves

almost the best resolution among four image restoration methods except for VEEV dataset. VEEV dataset is composed of ice and virus particles, in which the structure information is limited. Thus SC-Net may have problem in sparsity extraction while Topaz is trained on large dataset and has sufficient feature information. However, SC-Net can preserve more detailed membrane structures without introducing grid artifacts. Figure 9 presents $\text{FSC}_{e/o}^{-1}$ curves for Centriole and Vescicle, which shows that our SC-Net can perform image restoration with higher self-consistency and introduce fewer false signal comparing with other methods.

5.7. Ablation studies

To further clarify the effectiveness of our model, we conducted ablation studies on up-sampling block, 2D filter of smoothed volume generation and loss function L_{smooth}. These experiments were conducted on the simulated datasets Adhesion Belt and Synapse.

Study on Up-sampling block (UPB). We tested the original SC-Net (i.e., with UPB) and the SC-Net without UPB (i.e., non-UPB). Figure 10 shows the visual comparison and Table 4 shows the quantitative analysis of the output from these two models. From Figure 10 we can find that the results of SC-Net with UPB contain more detailed structure than the one without UPB. And From Table 4 we can find that the PSNR value of SC-Net without UPB has a drop of 0.77 dB for Adhesion Belt, and 5.08 dB for Synapse.

Study on Sparsity Representation Extractor. Here, 2D Topaz (N2N) Filter is replaced by BM3D. Figure 11 shows the visual results and Table 5 shows the quantitative results on Adhesion Belt and Synapse data trained with the tested two models. Results show that our SC-Net can still generate a result with noise smoothness despite the filter has changed. That is, our model is robust to different projection filters, which is a verification of SC-Net's model stability.

Study on L_{smooth}. We trained and tested SC-Net with a complete loss function in Eq. 8 and with a loss function excluding L_{smooth} (i.e., Non-L_{smooth}). Results show that L_{smooth} can provide strong improvement in noise removal when the training data is not sufficient. Complete results are available in Supplementary Material S5.3.

Table 4: PSNR(dB)/SSIM results for ablation study on Up-Sampling Block (UPB) under the noise with σ=15.

Dataset	Noisy	Non-UPB	With UPB
Synapse	26.02/0.316	22.47/0.868	**27.55/0.916**
Belt	23.44/0.282	24.91/0.881	**25.68/0.899**

Table 5: PSNR(dB)/SSIM results for ablation study on BM3D and Topaz Filter under the noise intensity σ=10.

Dataset	Noisy	BM3D-Filter	Topaz-Filter
Synapse	27.79/0.503	28.10/**0.948**	**30.03**/0.932
Belt	22.76/0.423	**33.30**/0.826	28.61/**0.914**

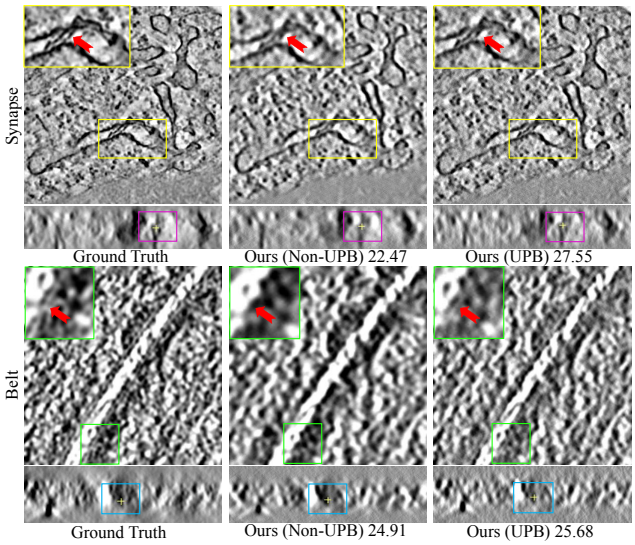


Figure 10: Visual comparisons between non-UPB SC-Net and SC-Net. (AWGN: $\sigma = 15$, metric: PSNR (dB)).

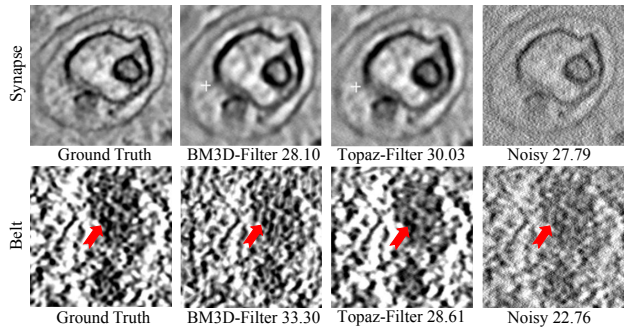


Figure 11: Visual comparisons between the results of SC-Net with smoothed volume filtered by BM3D and Topaz (AWGN: $\sigma = 10$, metric: PSNR (dB)).

6. Conclusion

In this article, we proposed the SC-Net, a novel self-supervised network for cryo-ET volumetric image restoration from single noisy data. SC-Net can make a balance between structure preservation and noise suppression to produce better image restoration. Comprehensive results proved that SC-Net could produce a strong enhancement for a single very noisy cryo-ET volumetric data, which is much better than Noise2Void and with a competitive performance comparing with Topaz.

Acknowledgement

This research is supported by the National Key Research and Development Program of China (No. 2017YFA0504702 and 2020YFA0712401), the NSFC projects Grants (61932018, 62072280, and 62072441), Beijing Municipal Natural Science Foundation Grant (No. L182053).

References

[1] Joshua Batson and Loic Royer. Noise2self: Blind denoising by self-supervision. *Proceedings of the 36th International Conference on Machine Learning*, 2019. 3

[2] Tristan Bepler, Kotaro Kelley, Alex J Noble, and Bonnie Berger. Topaz-denoise: general deep denoising models for cryoem and cryoet. *Nature communications*, 11(1):1–12, 2020. 3, 4, 5

[3] Antoni Buades, Bartomeu Coll, and J-M Morel. A non-local algorithm for image denoising. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 60–65. IEEE, 2005. 2

[4] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on image processing*, 16(8):2080–2095, 2007. 2

[5] Yuchen Deng, Yu Chen, Yan Zhang, Shengliu Wang, Fa Zhang, and Fei Sun. Icon: 3d reconstruction with 'missing-information' restoration in biological electron tomography. *Journal of Structural Biology*, 195(1):100 – 112, 2016. 2

[6] Jose-Jesus Fernandez. Computational methods for electron tomography. *Micron*, 43(10):1010 – 1030, 2012. 1

[7] Jose-Jesus Fernández and Sam Li. An improved algorithm for anisotropic nonlinear diffusion for denoising cryotomograms. *Journal of Structural Biology*, 144(1-2):152 – 161, 2003. 2

[8] Achilleas S. Frangakis and Reiner Hegerl. Noise reduction in electron tomographic reconstructions using nonlinear anisotropic diffusion. *Journal of Structural Biology*, 135(3):239 – 250, 2001. 2

[9] J. Frank, B. F. McEwen, M. Radermacher, J. N. Turner, and C. L. Rieder. Three-dimensional tomographic reconstruction in high voltage electron microscopy. *Journal of Electron Microscopy Technique*, 6(2):193–205, 1987. 1

[10] Shi Guo, Zifei Yan, Kai Zhang, Wangmeng Zuo, and Lei Zhang. Toward convolutional blind denoising of real photographs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[11] Renmin Han, Xiaohua Wan, Zihao Wang, Yu Hao, Jingrong Zhang, Yu Chen, Xin Gao, Zhiyong Liu, Fei Ren, Fei Sun, et al. Autom: a novel automatic platform for electron tomography reconstruction. *Journal of structural biology*, 199(3):196–208, 2017. 5

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 2

[13] Viren Jain and H. Sebastian Seung. Natural image denoising with convolutional networks. In *Proceedings of the 21st International Conference on Neural Information Processing Systems*, page 769–776, 2008. 2

[14] Wen Jiang, Matthew L Baker, Qiu Wu, Chandrajit Bajaj, and Wah Chiu. Applications of a bilateral denoising filter in biological electron microscopy. *Journal of Structural Biology*, 144(1-2):114 – 122, 2003. 2

[15] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR*, 2015. 4

[16] James R Kremer, David N Mastronarde, and J Richard McIntosh. Computer visualization of three-dimensional image data using imod. *Journal of structural biology*, 116(1):71–76, 1996. 5, 6

[17] Alexander Krull, Tim-Oliver Buchholz, and Florian Jug. Noise2void - learning denoising from single noisy images. In *CVPR*, June 2019. 1, 3

[18] Alexander Krull, Tomáš Vičar, Mangal Prakash, Manan Lalit, and Florian Jug. Probabilistic noise2void: Unsupervised content-aware denoising. *Frontiers in Computer Science*, 2020. 3

[19] Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. Noise2Noise: Learning image restoration without clean data. In *ICML*, pages 2965–2974, 2018. 1, 3, 4

[20] M. Maggioni, V. Katkovnik, K. Egiazarian, and A. Foi. Nonlocal transform-domain filter for volumetric data denoising and reconstruction. *IEEE Transactions on Image Processing*, 22(1):119–133, 2013. 2, 4

[21] Mauro Maiorca, Eric Hanssen, Edmund Kazmierczak, Bohumil Maco, Misha Kudryashev, Richard Hall, Harry Quiney, and Leann Tilley. Improving the quality of electron tomography image volumes using pre-reconstruction filtering. *Journal of Structural Biology*, 180(1):132–142, 2012. 1

[22] Xiao-Jiao Mao, Chunhua Shen, and Yubin Yang. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In *NIPS*, 2016. 2

[23] Nick Moran, Dan Schmidt, Yu Zhong, and Patrick Coady. Noisier2noise: Learning to denoise from unpaired noisy data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12064–12072, 2020. 3

[24] Ren Ng. Fourier slice photography. In *ACM SIGGRAPH 2005 Papers*, pages 735–744. 2005. 1

[25] Radosav S. Pantelic, Rosalba Rothnagel, Chang-Yi Huang, David Muller, David Woolford, Michael J. Landsberg, Alasdair McDowall, Bernard Pailthorpe, Paul R. Young, Jasmine Banks, Ben Hankamer, and Geoffery Ericksson. The discriminative bilateral filter: An enhanced denoising filter for electron microscopy data. *Journal of Structural Biology*, 155(3):395 – 408, 2006. 2

[26] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NIPS*, 12 2019. 4

[27] Yuhui Quan, Mingqin Chen, Tongyao Pang, and Hui Ji. Self2self with dropout: Learning self-supervised denoising from single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1890–1898, 2020. 3

[28] 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*, pages 234–241, 2015. 3

[29] Arne Stoschek and Reiner Hegerl. Denoising of electron tomographic reconstructions using multiscale transformations. *Journal of Structural Biology*, 120(3):257 – 265, 1997. 2

[30] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9446–9454, 2018. 3

[31] Peter van der Heide, Xiao-Ping Xu, Brad J. Marsh, Dorit Hanein, and Niels Volkmann. Efficient automatic noise reduction of electron tomographic reconstructions based on iterative median filtering. *Journal of Structural Biology*, 158(2):196 – 204, 2007. 2

[32] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[33] Rui Yan, Singanallur V Venkatakrishnan, Jun Liu, Charles A Bouman, and Wen Jiang. Mbir: A cryo-et 3d reconstruction method that effectively minimizes missing wedge artifacts and restores missing information. *Journal of structural biology*, 206(2):183–192, 2019. 2

[34] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, 2017. 2

[35] K. Zhang, W. Zuo, and L. Zhang. Ffdnet: Toward a fast and flexible solution for cnn-based image denoising. *IEEE Transactions on Image Processing*, 27(9):4608–4622, 2018. 2

[36] Yulun Zhang, Kunpeng Li, K. Li, B. Zhong, and Yun Fu. Residual non-local attention networks for image restoration. In *International Conference on Learning Representations (ICLR)*, 2019. 2