Supplementary Material for $\mathcal{J}$-Invariant Volume Shuffle for Self-Supervised Cryo-Electron Tomogram Denoising on Single Noisy Volume

S1. The Impact of 3D Pixel-Unshuffle/Shuffle on $\mathcal{J}$-invariance

In the context of Blind-Spot Networks (BSN) for self-supervised image denoising, ensuring $\mathcal{J}$-invariance is critical. $\mathcal{J}$-invariance means that the prediction for each pixel (or voxel in 3D a volumetric image) in the output image should not be influenced by the corresponding pixel in the input image. This property prevents the model from learning to replicate noisy inputs directly, thus avoiding identity mapping and ensuring effective denoising.

Generally, BSNs are composed of centrally masked convolution and dilated convolutions layers. Consider a BSN denoted as f, composed of d-dilated convolutions $f^{(l)}$ for all $l \in (1, L)$, with a kernel size of $3 \times 3 \times 3$. The function f can be expressed as:

$$f(x) = f^{(L)}(f^{(L-1)}(...f^{(1)}(f^{(0)}(x)))) \tag{S1}$$

where $f^{(0)}$ denotes a centrally masked convolutional layer using a $(2d-1) \times (2d-1) \times (2d-1)$ kernel, and x is the input noisy volume. The learned features for each convolutional layer l are represented as:

$$y^{(l)} = f^{(l)}(f^{(l-1)}(...f^{(1)}(f^{(0)}(x)))) \tag{S2}$$

Suppose $x_{i,j,k}$ is a voxel of a noisy volume x, where the coordinates i, j, k satisfies $i \mod q = 0$, $j \mod q = 0$, and $k \mod q = 0$. Here, q is the scale factor of 3D pixel-unshuffle. The dilated convolution will expose $x_{i,j,k}$'s features to its neighboring voxels. During 3D pixel-unshuffle, voxels $y^{(l)}_{m,n,o}$, where $i \leq m < i+q$, $j \leq n < j+q$, and $k \leq o < k+q$, are aligned to the channel axis of the same spatial position as $y^{(l)}_{i,j,k}$. Subsequent convolution operations would then expose $x_{i,j,k}$'s features to $y^{(l)}_{i,j,k}$, thus breaking the $\mathcal{J}$-invariance.

S2. Conditions for Maintaining $\mathcal{J}$-invariance

Maintaining $\mathcal{J}$-invariance in 3D self-supervised denoising networks, particularly when using volume-unshuffle operations, requires careful consideration of the relationship between the unshuffle volume size v and the dilation factor d in dilated convolution. Volume-unshuffle can only maintain $\mathcal{J}$-invariance when the volume size v is a multiple of the dilation factor d. This ensures that voxels with the same position as $x_{i,j,k}$ in each channel remain independent with $x_{i,j,k}$. The neighboring voxels of $x_{i,j,k}$ will be influenced after central masked convolution in the first layer, with the receptive field $\text{RF}(y^{(0)}, x_{i,j,k})$:

$$\text{RF}(y^{(0)}, x_{i,j,k}) = \{(i \pm (d-1), j \pm (d-1), k \pm (d-1)), \ldots, (i \pm 1, j \pm 1, k \pm 1)\} \tag{S3}$$

This means the central masked convolution creates a receptive field around the voxel $x_{i,j,k}$, affecting its neighboring voxels within a range determined by the dilation factor d. $\text{RF}(y^{(l)}, x_{i,j,k})$ indicates the receptive field of $x_{i,j,k}$ in $y^{(l)}$. The sequential, dilated convolution expand this receptive field:

$$\text{RF}(y^{(l)}, x_{i,j,k}) = \bigcup_{n_{\{1,2,3\}} \in \{-d,0,d\}} \{(i'+n_1, j'+n_2, k'+n_3) \mid (i', j', k') \in \text{RF}(y^{(l-1)}, x_{i,j,k})\} \tag{S4}$$

From Eq. S3, S4, we can infer that there still are voxels not influenced by $x_{i,j,k}$ with a period of d. When we apply volume-unshuffle to downsample $y^{(l)}$, we should keep these unaffected voxels staying at the same spatial position as $x_{i,j,k}$ in each channel. To facilitate clarity, we only focus on one spatial axis location as follow.

$$\text{Volume-Unshuffle}(y^{(l)}_i, v) = v \left\lfloor \frac{i}{v^3} \right\rfloor + (i \mod v) \tag{S5}$$

The voxels x_{i^*,j^*,k^*} that have same spatial position with $x_{i,j,k}$ will satisfy:

$$v \left\lfloor \frac{i^*}{v^3} \right\rfloor + (i^* \mod v) = v \left\lfloor \frac{i}{v^3} \right\rfloor + (i \mod v) \tag{S6}$$

$$|v(\left\lfloor \frac{i^*}{v^3} \right\rfloor - \left\lfloor \frac{i}{v^3} \right\rfloor)| = |(i - i^*) \bmod v| \tag{S7}$$

Judging from Eq. S7, the left side is a multiple of v, while the right side is the remainder when divided by v, which means that it must be smaller than v. For the equation to hold, both sides must be 0, meaning the positional difference between i^* and i must equal a multiple of v from i. Therefore, the independence on $x_{i,j,k}$ along the channel axis can be guaranteed with the unshuffle volume size v equals to the dilation factor d.

S3. Sources of Real Datasets

- The first dataset G. hansenii bio9-2 can be downloaded from CryoET Data Portal (https://cryoetdataportal.czscience.com/). The G. hansenii bio9-2 dataset is a tilt series of 61 projections ranging from $-60°$ to $+60°$ at $2°$ intervals. Each tilt image measures 960×928 pixels with 5.41 Å/pixel, acquired using a TFS Krios microscope with a Gatan K3 camera.
- The second dataset Vesicle is provided by the Institute of Biophysics, Chinese Academy of Sciences. The Vesicle dataset consists of 120 projections ranging from $-59°$ to $+60°$ at $1°$ intervals. Each tilt image measures 1024×1024 pixels with 8 Å/pixel, acquired using a TFS Talos Arctica microscope with a Falcon II camera.
- The third dataset Escherichia phage T4 is downloaded from EMDB (https://www.ebi.ac.uk/emdb/). The Escherichia phage T4 dataset contains 41 projections ranging from $-60°$ to $+60°$ at $3°$ intervals. Each tilt image measures 1024×1024 pixels with 1.558 Å/pixel, acquired using a FEI Tecnai F20 microscope with a Dectris Arina 4D-STEM detector.
- The fourth dataset Centriole can be downloaded from the IMOD tutorial (http://bio3d.colorado.edu/imod/files/tutorialData-1K.tar.gz).he Centriole dataset is a tilt series of 64 projections ranging from $-61.0°$ to $+65.0°$ at $2°$ at intervals. Each tilt image measures 1024×1024 pixels with 10.1 Å/pixel, acquired using an FEI TF39 microscope with a Gatan camera.

S4. Preparation of Simulated Dataset

As shown in Figure S1, the workflow for preparing a simulated Cryo-ET dataset involves introducing Additive White Gaussian Noise (AWGN) to noise-free projections, followed by the reconstruction of the resulting noisy data.

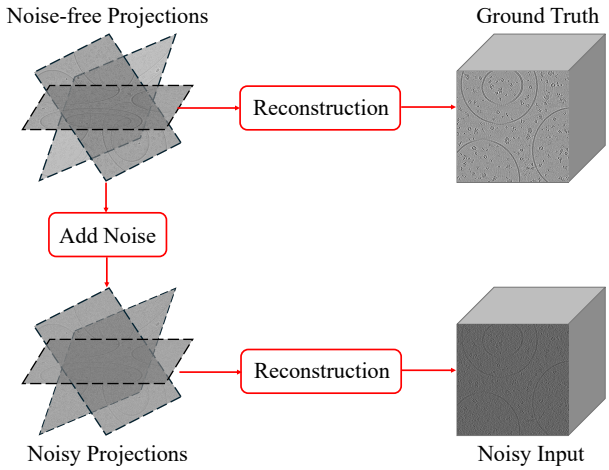

Figure S1: The workflow to prepare a simulated dataset with AWGN adding to noise-free projections.

Noise-free projections represent the ideal, high-quality data without any interference from noise. These noise-free projections undergo a reconstruction process to generate a 3D volume, referred to as the "Ground Truth". To simulate realistic imaging conditions, AWGN is introduced into the noise-free projections. This step replicates the inherent noise encountered

in actual Cryo-ET data acquisition, arising from sources such as electronic noise, environmental fluctuations, and the limitations of the imaging apparatus. Subsequently, these noisy projections are reconstructed to generate 3D volumetric images termed "Noisy Input". Figure S2 visualizes the reconstructed examples with noise intensity of σ = 0.2.

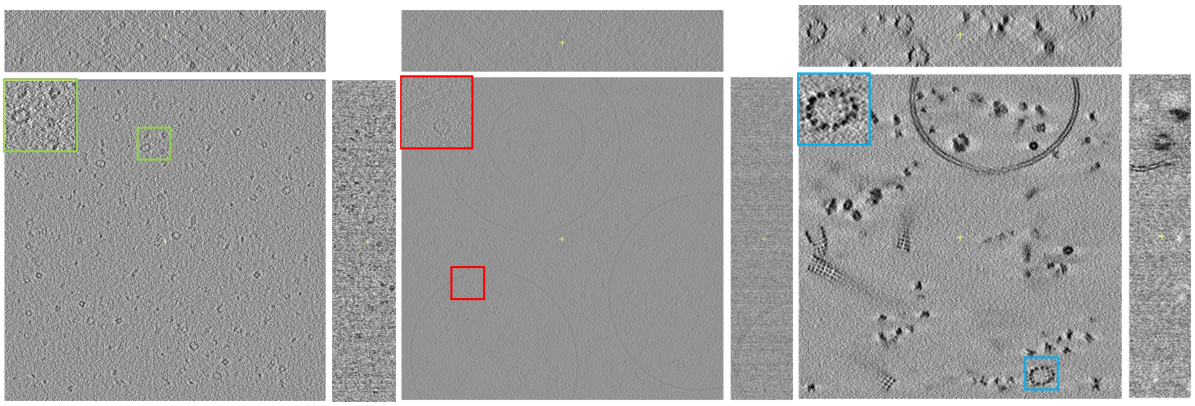

Figure S2: Examples of the noisy reconstructed volumetric data (metrics: PSNR (dB)/SSIM), in which the volumes are reconstructed from projections with AWGN (σ=0.2).

S5. Experiment Evaluation Metrics

S5.1. Simulated Dataset Experiments

In our simulated experiment, we utilize PSNR and SSIM as the primary metrics for evaluation. The mathematical formulations for PSNR and SSIM are provided in Eq. S8 and Eq. S9, respectively.

$$\text{PSNR}(\widetilde{V}, V^g) = 20 \log_{10} \left(\frac{\text{MAX}_I}{\text{MSE}} \right) \tag{S8}$$

$$\text{SSIM}(\widetilde{V}, V^g) = \frac{(2\mu_{\widetilde{V}}\mu_{V^g} + C_1)(2\sigma_{\widetilde{V}V^g} + C_2)}{(\mu_{\widetilde{V}}^2 + \mu_{V^g}^2 + C_1)(\sigma_{\widetilde{V}}^2 + \sigma_{V^g}^2 + C_2)} \tag{S9}$$

In the equations above, $\widetilde{V}$ denotes the denoised output image, while V^g represents the ground truth image. For our experiments, the pixel values of each image are normalized to the range $[0, 1]$. Consequently, the parameter MAX_I in Eq. S8 is set to 1. For Eq. S9, we use the constants $C_1 = (K_1 \cdot L)^2$ and $C_2 = (K_2 \cdot L)^2$, where $K_1 = 0.01$ and $K_2 = 0.03$. L represents the dynamic range of pixel values, which is set to 1 in our experiments due to normalization.

S5.2. Real Data Experiments

For real-world dataset without ground truth, we adopt a cross-validation metric called $\text{FSC}_{e/o}$ to assess the resolution of cryo-ET volumes. The Fourier shell correlation (FSC) between two tomographic volumes calculated from even and odd projections is defined as $\text{FSC}_{e/o}$. The mathematical formulation of FSC is

$$\text{FSC}(r) = \frac{\sum_{r_i \in r} F_1(r_i) \cdot F_2(r_i)^*}{\sqrt{\sum_{r_i \in r} |F_1(r_i)|^2 \sum_{r_i \in r} |F_2(r_i)|^2}} \tag{S10}$$

where F_1 is complex factor for volume 1, F_2^* conjugate of the structure factor for volume 2, and r_i is an individual voxel element at radius r. Assuming that the SNR in each tomographic volume from a half reconstruction has half signals of that in complete reconstruction, $\text{FSC}_{e/o}$ is calculated as:

$$\text{FSC}_{e/o}(r) = \frac{2\text{FSC}(r)}{\text{FSC}(r) + 1} \tag{S11}$$

S6. Edge Representation Enhancer.

Designed around the Kirsch operator, this edge representation enhancer excels in extracting detailed edge and contour information by leveraging multi-directional edge detection. The Kirsch operator takes a single kernel mask and rotates it in 45-degree increments through all 8 compass directions: North (N), Northwest (NW), West (W), Southwest (SW), South (S), Southeast (SE), East (E), and Northeast (NE). The edge magnitude of the Kirsch operator is calculated as the maximum magnitude across all directions:

$$h_{x,y,z} = \max_{n=1,\ldots,8} \sum_{i=-1}^{1} \sum_{j=-1}^{1} \sum_{k=-1}^{1} g_{ijk}^{(n)} \cdot f_{x+i,y+j,z+k} \tag{S12}$$

where $h_{x,y,z}$ is the edge magnitude at position (x, y, z), $g_{ijk}^{(n)}$ represents the Kirsch kernel for direction n, and $f_{x+i,y+j,z+k}$ is the voxel value at position $(x + i, y + j, z + k)$ in the 3D image.

The final edge intensity map $D(x, y, z)$ is obtained by applying thresholding operation, setting all pixels below a certain threshold T to 0. This effectively suppresses random edges caused.

$$D(x, y, z) = \begin{cases} h_{x,y,z} & \text{if } h_{x,y,z} \geq T \\ 0 & \text{if } h_{x,y,z} < T \end{cases} \tag{S13}$$

We conducted experiments to assess the impact of the edge representation enhancer in our denoising framework by comparing the Kirsch operator with the Sobel operator. Figure S3 reveals that the Kirsch operator excels at preserving edge details and demonstrates greater robustness against noise. Quantitative results in Table S1 corroborate these findings, with the Kirsch operator consistently delivering higher PSNR and SSIM values than the Sobel operator. This underscores the effectiveness of the Kirsch operator in capturing and enhancing edge information, thereby significantly improving the quality of the denoised volumetric images.

Table S1: PSNR/SSIM results for ablation study on Kirsch and sobel operator.(AWGN:σ=0.15).

Dataset	Noisy	Sobel	Kirsch
SHREC20	20.63/0.440	35.54/0.941	**36.18/0.952**
PolNet	23.67/0.448	36.92/0.934	**37.06/0.953**

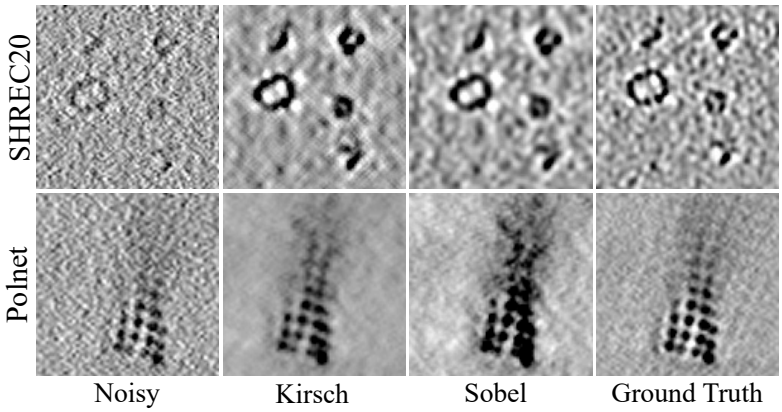

Figure S3: Visual results of study on edge representation enhancer.

S7. Visual Result of the Experiment on Simulated Data

To comprehensively analyze the results of experiments on simulated data, we provide additional visual results for each method. Figure S4 shows the visual comparison of the denoised tomogram, where the displayed images are selected from the middle slice of the tomograms along the direction of x-, y- and z-axis. Judging from Figure S4, our method shows superior performance in preserving structural details and reducing noise, with the tomogram exhibiting fewer artifacts compared to those processed by other methods.

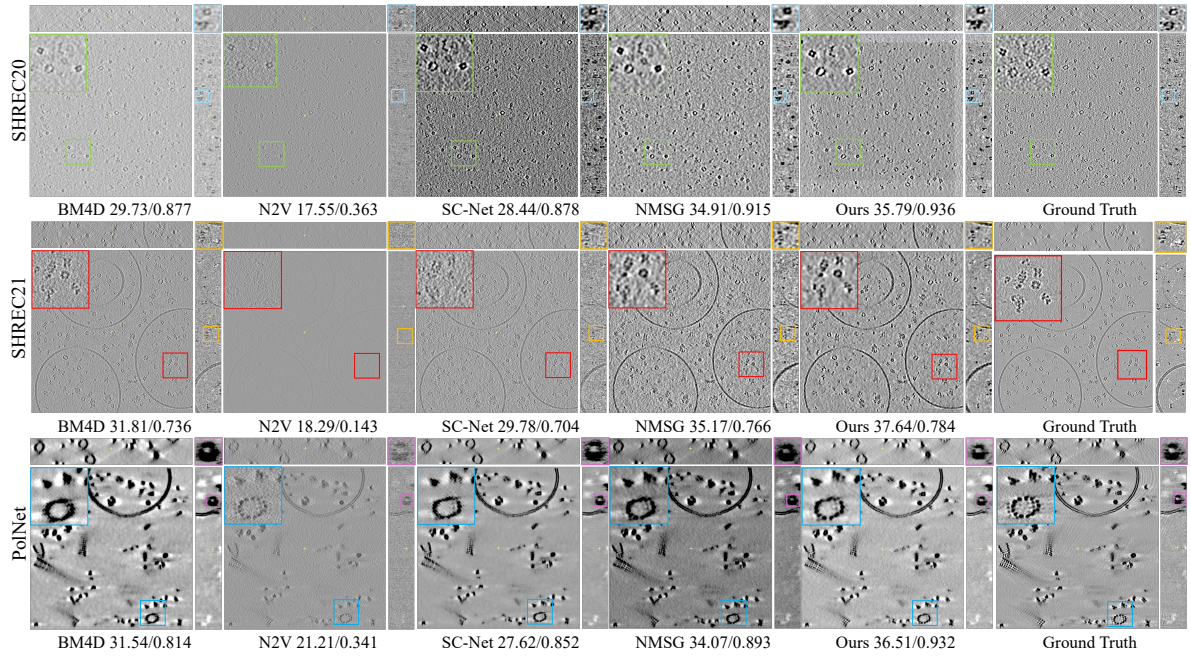

Figure S4: Visual results of the simulated data with AWGN ($\sigma = 0.2$) shown in 3D space (metrics: PSNR(dB)/SSIM).