A supervised multi-task framework for joint cryo-ET restoration enabled by generative physical simulation

Supplementary Material

Content

This supplementary material provides the following contents:

- Details about the network architecture (Section A).
- Evaluation metrics for the simulated dataset (Section B).
- Selected areas for CNR and ENL (Section C).
- Additional baseline comparisons (Section D).
- Ablation studies on noise models (Section E).
- Ablation studies on λ_{recon} (Section F).

A. Network Architecture

Noise synthesizer. The noise synthesizer is implemented based on the WGAN-GP framework using convolutional neural networks. The generator consists of five transposed convolutional blocks, each containing a 2D transposed convolutional layer, BatchNorm, and ReLU. The output layer is a 2D transposed convolutional layer with a stride of 2. The discriminator has five convolutional blocks, each consisting of a 2D convolutional layer, BatchNorm, and Leaky ReLU with slope k of 0.2. The output layer is a 2D convolutional layer with a stride of 1. In the trained GAN, the generator directly synthesizes 2D noise patches matching the size of the extracted noise patches. These synthetic 2D noise patches are stacked into 3D noise volumes for synthesizing noisy input volumes.

Restoration network. The restoration network is based on the U-Net architecture [1]. Before the synthetic noisy input volume is fed into the encoders, it first passes through an upsampling block implemented with nearest neighbor interpolation to enhance structural information during training. The encoder adopts five convolutional blocks, each consisting of a max pooling layer, 3D convolutional layer, and Leaky ReLU. The decoder uses five deconvolutional blocks, each containing an interpolation layer, two 3D convolutional layers, and Leaky ReLU. To better recover detailed information, skip connections are used between corresponding convolutional and deconvolutional blocks at matching spatial resolutions.

B. Evaluation metrics for simulated dataset

PSNR and SSIM are selected as measures for our evaluations in the simulated experiment. Eq.1 and Eq.2 give the mathematical formulations of PSNR and SSIM, respectively.

$$\text{PSNR}\left(\tilde{\boldsymbol{V}}, \boldsymbol{V}^g\right) = 20\log_{10}\left(\frac{\text{MAX}_I}{\text{MSE}}\right) \tag{1}$$

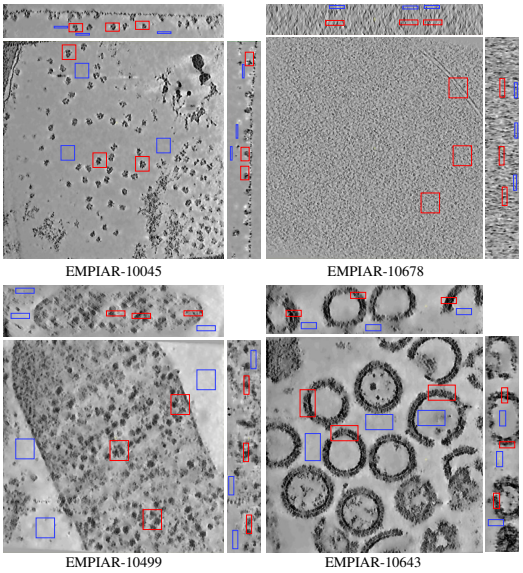

Figure 1. Demonstrations of the selected areas in realistic datasets. Red areas refer to foreground areas, blue areas refer to background areas. For each result, the main panel shows a representative x–y slice, with the corresponding x–z slice (top) and y–z slice (right).

$$\text{SSIM}\left(\tilde{\boldsymbol{V}}, \boldsymbol{V}^g\right) = \frac{(2\mu_{\tilde{\boldsymbol{V}}}\mu_{\boldsymbol{V}^g} + C_1)(2\sigma_{\tilde{\boldsymbol{V}}\boldsymbol{V}^g} + C_2)}{\left(\mu^2_{\tilde{\boldsymbol{V}}} + \mu^2_{\boldsymbol{V}^g} + C_1\right)\left(\sigma^2_{\tilde{\boldsymbol{V}}} + \sigma^2_{\boldsymbol{V}^g} + C_2\right)} \tag{2}$$

In the equations above, $\tilde{\boldsymbol{V}}$ denotes the output image, while $\boldsymbol{V}^g$ refers to the ground truth image. In our experiments, the pixel values of all images are normalized to the range $[0, 1]$, which means the parameter MAX_I in Eq.1 is set to 1. For Eq.2, based on the study in [2], $C1 = (K_1 * L)^2$ and $C2 = (K_2 * L)^2$, where $K_1 = 0.01$, $K_2 = 0.03$, and $L = 255$ for 8-bit images. Since the maximum pixel value is 1 due to our normalization, L is set to 1 in our experiments.

C. Selected areas for CNR and ENL

The computation of CNR and ENL requires separately extracting areas with foreground and background. In our experiments, we select 3 areas containing foreground (s_1, s_2, and s_3) and 3 areas only containing background (b_1, b_2, and b_3). Figure 1 shows the selected areas pointed out on the tomograms. Table 1-4 presents the range of coordinates for selected areas.

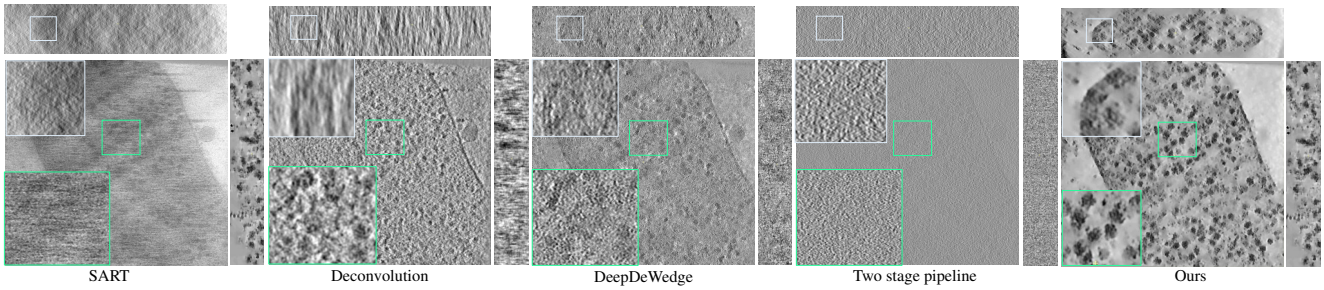

Figure 2. Visual comparison of baseline methods on the EMPIAR-10499 dataset.

Table 1. Coordinates of the selected areas for CNR and ENL (Dataset: EMPIAR-10045)

areas	s_1	s_2	s_3	b_1	b_2	b_3
z-min	50	50	50	50	50	50
z-max	60	60	60	60	60	60
y-min	400	390	830	800	460	480
y-max	440	430	870	840	500	520
x-min	400	590	270	330	280	620
x-max	430	630	300	370	320	660

Table 2. Coordinates of the selected areas for CNR and ENL (Dataset: EMPIAR-10499)

areas	s_1	s_2	s_3	b_1	b_2	b_3
z-min	190	190	190	190	190	190
z-max	200	200	200	200	200	200
y-min	420	110	580	50	400	800
y-max	520	210	680	150	500	900
x-min	460	550	560	120	30	800
x-max	560	650	660	220	130	900

Table 3. Coordinates of the selected areas for CNR and ENL (Dataset: EMPIAR-10678)

areas	s_1	s_2	s_3	b_1	b_2	b_3
z-min	70	70	70	130	130	130
z-max	75	75	75	135	135	135
y-min	1040	920	560	1040	920	560
y-max	1090	970	610	1090	970	610
x-min	1400	1360	1450	1400	1360	1450
x-max	1450	1410	1500	1450	1410	1500

Table 4. Coordinates of the selected areas for CNR and ENL (Dataset: EMPIAR-10643)

areas	s_1	s_2	s_3	b_1	b_2	b_3
z-min	285	285	285	285	285	285
z-max	295	295	295	295	295	295
y-min	710	670	730	400	880	500
y-max	760	720	760	450	930	550
x-min	430	170	750	750	270	180
x-max	450	200	800	800	320	230

D. Additional Baseline Comparisons

To further validate the effectiveness of the proposed method, we compare it against several additional baselines, including SART (iterative reconstruction), Deconvolution, and DeepDeWedge. In addition, we implement a two-stage pipeline baseline, where the tilt series is first denoised and the tomogram is subsequently reconstructed using WBP.

As shown in Table 5, our method achieves the best CNR and ENL metrics across both the EMPIAR-10045 and EMPIAR-10499 datasets. The inferior performance of the two-stage design can be attributed to the inherently lower SNR of tilt series compared with reconstructed tomograms, which makes projection-domain denoising more difficult and more prone to suppressing weak macromolecular signals. Visual comparisons of these baselines on the EMPIAR-10499 dataset are provided in Figure 2.

Table 5. Comparison to additional baselines (metrics: CNR/ENL). Larger values indicate better performance.

Dataset	EMPIAR-10045	EMPIAR-10499
SART	0.168 / 72.674	0.020 / 55.093
Deconvolution	0.257 / 20.342	0.188 / 17.104
DeepDeWedge	0.033 / 22.27	0.353 / 128.602
Two-stage pipeline	0.286 / 144.31	0.049 / 13.458
Ours	**0.506 / 342.288**	**1.637 / 162.764**

E. Ablation Studies on Noise Models

We compared our WGAN-GP noise model with Gaussian noise under the same pipeline. The quantitative results are shown in Table 6, and the visual comparison on the EMPIAR-10499 dataset is shown in Figure 3. The results show that Gaussian noise with different target SNR values cannot well capture the noise characteristics in cryo-ET data, and the corresponding restorations still contain noticeable artifacts or degraded structures. By comparison, our WGAN-GP noise model produces clearer structures and cleaner backgrounds, showing better visual quality overall. These results are consistent with the quantitative results in Table 6.

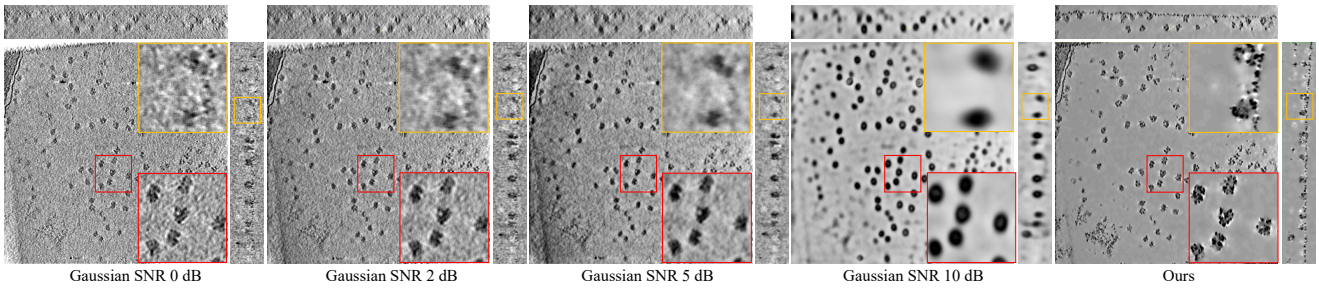

Figure 3. Visual comparison of different noise models on the EMPIAR-10045 dataset.

Table 6. Ablation studies on noise types. Gaussian noise is i.i.d. and energy matched per tilt to target SNR (dB) values $\{0, 2, 5, 10\}$ (metrics: CNR/ENL).

Dataset	EMPIAR-10045	EMPIAR-10499
Gaussian SNR 0 dB	0.388 / 139.478	0.318 / 83.220
Gaussian SNR 2 dB	0.334 / 99.288	0.353 / 69.047
Gaussian SNR 5 dB	0.324 / 71.786	0.3514 / 50.312
Gaussian SNR 10 dB	0.365 / 92.854	0.492 / 47.746
WGAN-GP (ours)	**0.506 / 342.288**	**1.637 / 162.764**

F. Ablation Studies on λ_{recon}

We perform an ablation study on λ_{recon}, and the quantitative results are summarized in Table 7. A visual comparison on the EMPIAR-10499 dataset is further provided in Figure 4. The weight λ_{recon} mainly determines the balance between noise suppression and missing wedge restoration. A too small or too large λ_{recon} both introduce visible artifacts, while $\lambda_{\text{recon}} = 0.1$ gives a balance according to both the quantitative and visual results.

Table 7. Ablation studies on λ_{recon} sensitivity (metrics: CNR/ENL).

Dataset	EMPIAR-10045	EMPIAR-10499
λ_{recon}=0.0	0.446 / 227.332	1.045 / 48.432
λ_{recon}=0.1 (ours)	0.506 / **342.288**	**1.637 / 162.764**
λ_{recon}=0.2	0.570 / 337.703	1.587 / 116.165
λ_{recon}=1.0	**0.704** / 210.108	1.417 / 109.931

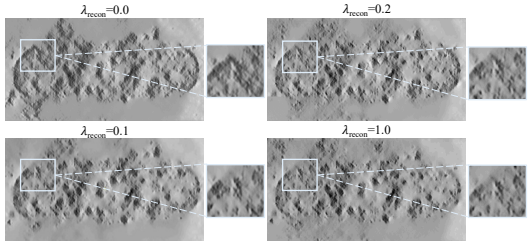

Figure 4. Visual comparison of different λ_{recon} settings on the EMPIAR-10499 dataset. The displayed images are x–z slices, shown to highlight the artifact differences under different parameter settings.

References

[1] 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.

[2] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004.