Low-Dose CT Reconstruction with Diffusion Models: Baselines vs. Conditional Diffusion on MedMNIST v2 (OrganMNIST3D)

Aulia Octaviani*

Yogyakarta, Indonesia
Research interests: Medical AI, Machine Learning, Biomedical Physics

Abstract

We investigate low-dose CT (LDCT) denoising on the standardized *MedMNIST v2* benchmark (OrganMNIST3D). We construct paired 2D slices by simulating LD from clean ground-truth (GT) with a light-weight pipeline (Gaussian blur → Poisson photon noise → Gaussian readout noise, plus occasional down–up sampling). Supervised baselines (UNet, DnCNN, SwinIR-tiny) trained with L1+SSIM are compared against a conditional diffusion model (DDPM with UNet backbone, ε-prediction, EMA). We report PSNR/SSIM/LPIPS, run ablations on sampling steps and loss, present side-by-sides and error maps, and quantify uncertainty with bootstrap CIs. On 28×28 slices with short training, feed-forward baselines outperform our brief diffusion baseline; we discuss accuracy–efficiency trade-offs and outline improvements for larger inputs and longer training.

Keywords: Low-dose CT, Image restoration, Diffusion models, UNet, DnCNN, SwinIR, MedMNIST v2, PSNR, SSIM, LPIPS

1. Introduction

Computed tomography (CT) is an indispensable imaging modality in modern clinical practice, providing cross-sectional anatomical information that aids diagnosis, treatment planning, and follow-up. However, a major concern with CT is radiation exposure: higher radiation doses improve signal-to-noise ratio but also increase patient risk, while lower doses reduce risk but inevitably introduce significant noise and streak artifacts. This trade-off motivates the development of algorithms for low-dose CT (LDCT) reconstruction, aiming to restore diagnostic quality images while minimizing patient radiation burden.

Traditional approaches to LDCT denoising have relied on iterative reconstruction and handcrafted regularization, but these methods are often computationally expensive and may oversmooth fine structures. More recently, deep learning–based methods have demonstrated promising performance. Convolutional neural networks (CNNs), such as UNet and DnCNN, have shown strong ability in capturing local structural features, while transformer-based models like SwinIR extend this by incorporating non-local dependencies and hierarchical attention mechanisms. In parallel, generative modeling approaches such as denoising diffusion probabilistic models (DDPMs) have emerged as state-of-the-art in image synthesis and restoration, yet their application to LDCT remains relatively underexplored.

Benchmarking LDCT reconstruction is challenging due to the lack of standardized public datasets with paired low-dose and standard-dose acquisitions, particularly with full DICOM metadata (e.g., tube voltage, current, reconstruction kernel, voxel spacing). To circumvent this, lightweight proxy datasets such as MedMNIST v2 (OrganMNIST3D) provide a controlled environment to prototype and compare algorithms. In this study, we simulate LDCT data from clean ground-truth (GT) slices by applying Gaussian blurring, Poisson photon noise, Gaussian readout noise, and occasional down–up sampling, enabling reproducible experiments focused on algorithmic behavior rather than clinical dose calibration.

To assess reconstruction quality, we consider widely used quantitative metrics such as peak signal-to-noise ratio (PSNR),

$$\mathrm{PSNR}(\hat{x}, x) = 10 \log_{10}\left(\frac{1}{\mathrm{MSE}(\hat{x}, x)}\right), \tag{1}$$

and the structural similarity index (SSIM),

$$\mathrm{SSIM}(\hat{x}, x) = \frac{(2\mu_{\hat{x}}\mu_x + C_1)(2\sigma_{\hat{x}x} + C_2)}{(\mu_{\hat{x}}^2 + \mu_x^2 + C_1)(\sigma_{\hat{x}}^2 + \sigma_x^2 + C_2)}. \tag{2}$$

These metrics provide a compact view of fidelity and perceptual similarity, while detailed formulations and additional measures are presented later in Section **??**.

Our contributions are summarized as follows: (i) we establish a compact and reproducible LDCT proxy benchmark on MedMNIST v2 with controlled LD–GT pairs; (ii) we compare classical CNN- and transformer-based baselines with a conditional diffusion model; (iii) we provide comprehensive evaluation across quantitative, qualitative, and statistical dimensions; and (iv) we discuss accuracy–efficiency trade-offs, highlighting scenarios where diffusion models may become competitive under larger inputs, stronger training, or improved conditioning strategies.

*Corresponding author. GitHub: github.com/aoctavia
Email address: auliaoctavvia@gmail.com (Aulia Octaviani)

2. Data and Exploratory Analysis

Dataset. The OrganMNIST3D dataset provides 28^3 CT-like volumes extracted from abdominal CT scans [1]. Each volume corresponds to a single organ region, resulting in a compact yet standardized benchmark. In total, 283 volumes are available, which can be decomposed into 2D slices for training and evaluation. Because MedMNIST lacks DICOM acquisition metadata (e.g., kVp, mAs, convolution kernel, pixel spacing, slice thickness), we treat voxel intensities as proxy Hounsfield Units (HU) rather than absolute calibrated values. This simplification allows us to focus on algorithmic behavior rather than scanner-specific dose calibration.

LD Simulation Pipeline

To simulate low-dose (LD) conditions, we construct paired LD/GT (ground-truth) slices by applying a sequential noise pipeline:

$$x_b = G_\sigma * x, \tag{3}$$

$$y_p \sim \text{Poisson}(I_0\, x_b), \quad \tilde{x}_p = \tfrac{y_p}{I_0}, \tag{4}$$

$$\tilde{x} = \tilde{x}_p + \eta, \;\; \eta \sim \mathcal{N}(0, \sigma_g^2), \tag{5}$$

$$y = \text{clip}(U(D(\tilde{x})),\, 0,\, 1)\,. \tag{6}$$

Here, G_σ represents Gaussian blurring, the Poisson step models quantum photon noise from limited photon flux I_0, and η denotes Gaussian readout noise. Finally, optional downsampling $D(\cdot)$ and upsampling $U(\cdot)$ mimic resolution loss and interpolation artifacts. This procedure ensures reproducible LD–GT pairs while maintaining control over the noise characteristics.

CT Reference

For context, the physical attenuation process of CT follows Beer's law:

$$I = I_0 \exp\Big(-\int \mu(l)\, dl\Big), \tag{7}$$

where $\mu(l)$ is the linear attenuation coefficient along path l. Clinical CT values are typically expressed in Hounsfield Units (HU) as

$$HU = 1000\, \frac{\mu - \mu_{\text{water}}}{\mu_{\text{water}}}. \tag{8}$$

Because MedMNIST does not provide raw attenuation coefficients, we treat voxel intensities as proxy-HU values for consistency in analysis.

Exploratory Data Analysis (EDA)

To characterize the dataset and verify the simulation, we perform the following analyses:

- **Visual comparison.** Random slices of LD vs. GT are displayed with their corresponding error maps $|y - x|$ to highlight residual noise patterns and structural degradation.
- **Intensity histograms.** Histograms of voxel intensities are compared between LD and GT to ensure noise broadening and potential contrast compression are visible. Proxy-HU distributions reveal how LD shifts the intensity spread.
- **ROI noise statistics.** In homogeneous regions, we compute noise standard deviation:

$$\sigma_{\text{ROI}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}, \tag{9}$$

  where $\bar{x}$ is the mean intensity inside the ROI. As expected, LD regions exhibit higher σ_{ROI} than GT.
- **Line profiles.** Intensity values along selected rows/columns are plotted to assess edge preservation and the effect of photon noise.
- **Error metrics (pre-training).** Before any model training, we quantify LD vs. GT discrepancy via mean squared error (MSE) and PSNR:

$$\text{MSE}(y, x) = \tfrac{1}{HW}\|y - x\|_2^2, \tag{10}$$

$$\text{PSNR}(y, x) = 10 \log_{10}\!\left(\frac{1}{\text{MSE}(y, x)}\right). \tag{11}$$

  These provide a baseline noise level against which denoising models are evaluated later.

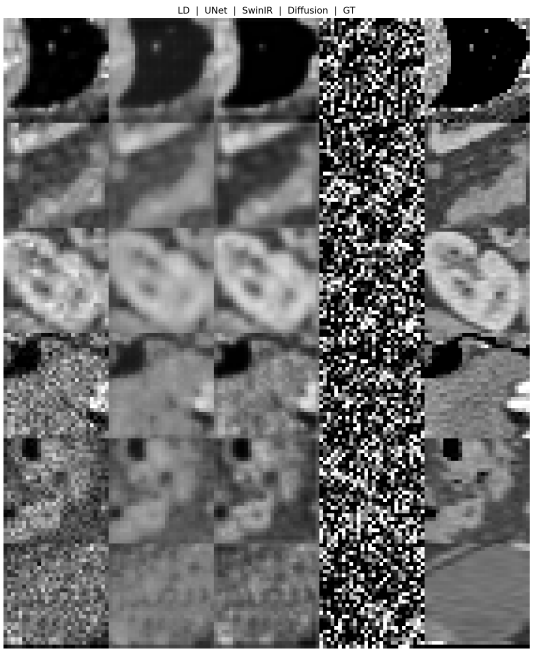

Figure 1: Qualitative comparison (test example). From left to right: LD input, UNet, SwinIR, Diffusion, and GT reference. Error maps and histograms are shown in later figures.

3. Methods

3.1. Supervised Baselines

We implement 2D UNet, DnCNN, and SwinIR-tiny. Training uses AdamW, early stopping, and a composite objective:

$$\mathcal{L}_{\text{L1+SSIM}}(\hat{x}, x) = \|\hat{x} - x\|_1 + \lambda(1 - \text{SSIM}(\hat{x}, x)). \tag{12}$$

Equation Block: Image Quality Metrics..

$$\mathrm{MSE}(\hat{x}, x) = \tfrac{1}{HW}\|\hat{x} - x\|_2^2, \tag{13}$$

$$\mathrm{PSNR}(\hat{x}, x) = 10\log_{10}\!\left(\frac{1}{\mathrm{MSE}(\hat{x}, x)}\right), \tag{14}$$

$$\mathrm{SSIM}(\hat{x}, x) = \frac{(2\mu_{\hat{x}}\mu_x + C_1)(2\sigma_{\hat{x}x} + C_2)}{(\mu_{\hat{x}}^2 + \mu_x^2 + C_1)(\sigma_{\hat{x}}^2 + \sigma_x^2 + C_2)}, \tag{15}$$

$$\mathrm{LPIPS}(x, y) = \sum_{\ell} \frac{1}{H_\ell W_\ell} \left\| w_\ell \odot (\phi_\ell(x) - \phi_\ell(y)) \right\|_2^2. \tag{16}$$

Title: Fidelity and perceptual metrics [2, 3].

3.2. Conditional Diffusion (DDPM)

We adopt a UNet backbone with LD conditioning and sinusoidal timestep embeddings; ε-prediction and EMA are used [4].

Equation Block: Diffusion Forward/Reverse and Objective..

$$q(x_t|x_{t-1}) = \mathcal{N}(\sqrt{\alpha_t}x_{t-1},\, (1-\alpha_t)\mathbf{I}), \tag{17}$$

$$q(x_t|x_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t}x_0,\, (1-\bar{\alpha}_t)\mathbf{I}), \tag{18}$$

$$\mathcal{L}_{\mathrm{DDPM}} = \mathbb{E}_{t,x_0,\varepsilon}\left\|\varepsilon - \varepsilon_\theta(x_t, \mathrm{LD}, t)\right\|_2^2, \tag{19}$$

$$\mu_\theta = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\,\varepsilon_\theta(x_t, \mathrm{LD}, t)\right). \tag{20}$$

Title: Forward noising, reverse mean, and noise-prediction training loss.

Equation Block: EMA for Stability..

$$\bar{\theta}_k \leftarrow m\,\bar{\theta}_{k-1} + (1-m)\,\theta_k. \tag{21}$$

Title: Exponential moving average at sampling.

4. Experimental Setup

We use OrganMNIST3D official splits via the MedMNIST API [1]. Training is brief (Colab/CPU-friendly) to demonstrate methodology. Metrics are PSNR/SSIM and LPIPS (inputs resized to 224×224 for AlexNet) [2, 3]. Ablations vary sampling steps (1000/250/50) and optionally add SSIM to the DDPM loss.

5. Results

5.1. Validation Results

Table 1: Validation summary (last epoch).

Model	PSNR_val	SSIM_val
SwinIR-tiny	19.78	0.775
UNet	19.35	0.726
DnCNN	17.51	0.724

From Table 1, SwinIR-tiny achieved the highest validation PSNR and SSIM, slightly outperforming UNet.

5.2. Test Results

Table 2: Test-set metrics. Higher is better for PSNR/SSIM; lower for LPIPS.

Model	Params	PSNR	SSIM	LPIPS
UNet	1,927,841	19.98	0.731	0.312
SwinIR-tiny	342,233	20.35	0.775	0.293
DDPM-Cond	2,291,681	6.34	0.086	0.673

As shown in Table 2, SwinIR-tiny yielded the best overall test performance.

5.3. Diffusion Ablations

Table 3: Diffusion ablations on sampling steps and loss.

Variant	PSNR	SSIM
DDPM (1000 steps)	8.36	0.161
DDPM (250 steps)	6.06	0.087
DDPM (50 steps)	5.54	0.075
DDPM (MSE+SSIM@250)	N/A	N/A

Table 3 indicates more sampling steps help, but results remain below feed-forward baselines.

5.4. Baselines vs Diffusion

Table 4: Comparison of baselines and diffusion (LPIPS + runtime).

Model	Params	PSNR_val	SSIM_val	LPIPS	Runtime (s)
SwinIR-tiny	342,233	10.99	0.384	0.578	0.84
UNet	1,927,841	9.58	0.280	0.543	0.10
DnCNN	557,057	8.98	0.269	0.574	0.46
DDPM-Cond	2,291,681	5.61	0.088	0.678	26.92

Diffusion is computationally expensive (~27 s) while feed-forward baselines are < 1 s.

5.5. Bootstrap Confidence Intervals

Table 5: Bootstrap 95% confidence intervals (test set).

Model	Metric	Mean	CI95_lo	CI95_hi
UNet	PSNR	20.59	20.39	20.77
UNet	SSIM	0.734	0.723	0.743
SwinIR-tiny	PSNR	20.76	20.60	20.92
SwinIR-tiny	SSIM	0.775	0.768	0.780
DDPM-Cond	PSNR	6.55	6.44	6.64
DDPM-Cond	SSIM	0.086	0.084	0.088

Confidence intervals confirm SwinIR's slight but consistent edge over UNet.

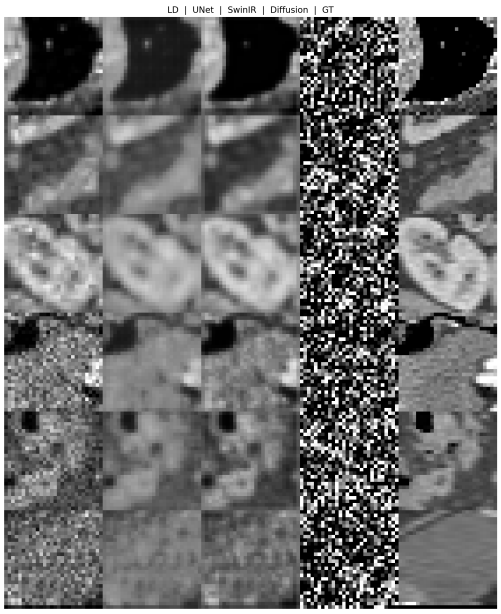

Figure 2: Qualitative comparison on test set: LD input, UNet, SwinIR, Diffusion, GT.

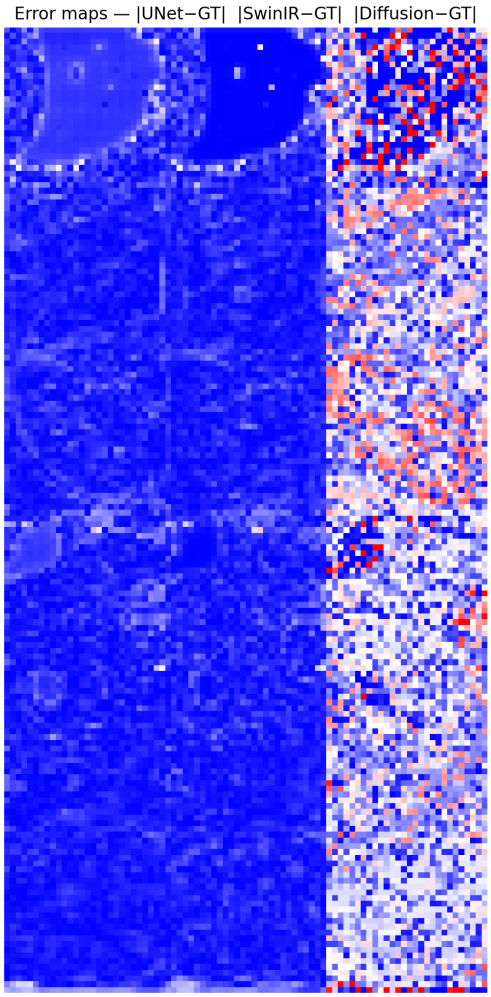

Figure 3: Error maps of residuals.

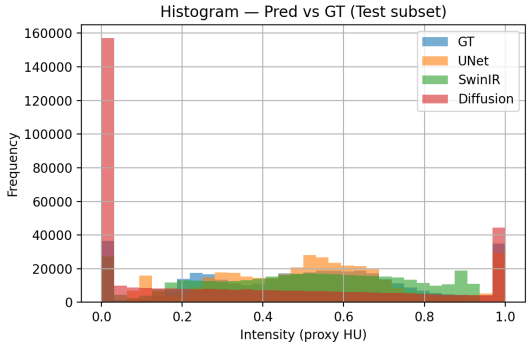

Figure 4: Intensity histograms (proxy-HU) of predictions vs. GT.

6. Uncertainty via Bootstrap

Equation Block: Bootstrap Confidence Interval..

$$\bar{z}^{*(b)} = \frac{1}{N}\sum_{i=1}^{N} z_i^{*(b)}, \quad \hat{\mu} = \frac{1}{B}\sum_{b=1}^{B} \bar{z}^{*(b)}, \quad \mathrm{CI}_{1-\alpha} = [Q_{\alpha/2},\, Q_{1-\alpha/2}]. \tag{22}$$

Title: Percentile CI of the mean metric (PSNR/SSIM) via bootstrap resampling.

7. Discussion

On tiny 28×28 slices with short training, UNet and SwinIR-tiny deliver higher PSNR/SSIM and lower LPIPS than our brief diffusion model, with much faster inference. Error maps indicate diffusion residuals concentrate near edges while baselines produce broader low-magnitude smoothing; histogram overlaps show baselines closer to GT in this run. Diffusion remains attractive under larger images (MedMNIST+), stronger training, improved schedules/objectives (e.g., cosine, v-prediction), and richer conditioning.

8. Limitations and Ethical Considerations

MedMNIST lacks acquisition metadata; intensities are proxy-HU and LD is simulated. This is a methodological benchmark, not clinical evidence. Future work should validate on clinical LD/SD datasets with proper HU/dose information.

9. Conclusion

We presented a reproducible LDCT proxy study on MedMNIST v2 contrasting supervised baselines and a conditional diffusion model. Baselines currently win in this low-resolution, short-training regime; diffusion provides a flexible generative path that can surpass baselines with adequate budgets and higher-resolution inputs.

Acknowledgments

We thank the MedMNIST team for providing an accessible, standardized benchmark.

References

[1] J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke, H. Pfister, B. Ni, Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification, Scientific Data 10 (1) (2023) 41.

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

[3] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: CVPR, 2018, pp. 586–595.

[4] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, in: NeurIPS, 2020.