Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Robust Evaluation of Low-Dose CT Reconstruction Through Generative and Discriminative Models

Jose M. Saavedra

jmsaavedrar@miuandes.cl

Universidad de Los Andes, Chile

Tomás Sotta

Universidad de Los Andes, Chile

Violeta Chang

Universidad de Santiago de Chile

Aline Xavier

Universidad de Santiago de Chile

Héctor Henríquez

Universidad de Los Andes, Chile

Jose Delpiano

Universidad de Los Andes, Chile

Camila Figueroa

Universidad de Chile

Joaquín Curimil

Universidad de Chile

Julio Serrano

Universidad de Santiago de Chile

Yessenia Orellana

Universidad de Los Andes, Chile

Article

Keywords:

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11652856/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A Robust Evaluation of Low-Dose CT Reconstruction Through Generative and Discriminative Models

Jose M. Saavedra[1*], Tomás de la Sotta[1], Violeta Chang[2], Aline Xavier[3], Héctor Henríquez[4], Jose Delpiano[1], Camila Figueroa[4], Joaquín Curimil[4], Julio Serrano[2], Yessenia Orellana[5]

[1]Facultad de Ingeniería y Ciencias Aplicadas, Universidad de los Andes, Chile, Mons. Álvaro del Portillo 12455, Santiago, 7550000, RM, Chile.
[2]Departamento de Ingeniería Informática, Universidad de Santiago de Chile, Av. Victor Jara 3659, Santiago, 9170022, RM, Chile.
[3]Facultad de Ingeniería, Universidad de Santiago de Chile, Av. Victor Jara 3659, Santiago, 9170022, RM, Chile.
[4]Facultad de Medicina, Universidad de los Andes, Chile, Mons. Álvaro del Portillo 12455, Las Condes, Santiago, 7550000, RM, Chile.
[5]Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile, Beauchef 851, Santiago, 8370459, RM, Chile.

*Corresponding author(s). E-mail(s): jmsaavedrar@miuandes.cl;
Contributing authors: tadelasotta@miuandes.cl; violeta.chang@usach.cl; aline.xavier@usach.cl; hhenriquez@miuandes.cl; jfdelpiano@miuandes.cl; cafigueroa@ug.uchile.cl; joaquin.curimil@ug.uchile.cl; julio.serrano@usach.cl; yorellana@miuandes.cl;

Abstract

Low-dose CT (LDCT) reconstruction is a critical problem in radiology that has a significant impact on public health. With the rise of deep learning, especially in generative models, we have seen a diverse range of methods for reconstructing LDCT images. However, existing LDCT reconstruction methods are often evaluated under inconsistent experimental settings, including different dataset splits, preprocessing pipelines, and noise-generation procedures, hindering fair comparisons across approaches. In addition, many studies do not explicitly assess

generalization across diverse datasets with different acquisition characteristics, limiting the practical relevance of the reported results. Therefore, we present a comprehensive fair study of modern generative models for LDCT reconstructions. Beyond effectiveness on a single dataset, our main goal is to assess the models' generalization across datasets. Moreover, we propose a new semantic segmentation-based metric, in addition to traditional image quality metrics, to evaluate the effectiveness of the reconstructed images in a medical task. Our results show that diffusion-based models with a small number of iterations outperform UNet and GAN, especially on out-of-distribution datasets. However, there are still challenges that require further study regarding effectiveness and efficiency.

Keywords: Computer tomography, diffusion models, generative models, low-dose CT reconstruction

1 Introduction

Radiological imaging has become an essential tool for disease diagnosis, progression monitoring, and image-guided therapeutic interventions. Diagnostic imaging modalities that use ionizing radiation are among the primary sources of radiation exposure in the general population, accounting for approximately 50% of the total ionizing radiation exposure (from both natural and artificial sources) [1]. In fact, over the past few decades, the number of radiological examinations has increased considerably, enabling faster and more accurate diagnoses and while reducing reliance on invasive high-risk procedures [1].

Computed tomography (CT) is one of the most widely used imaging modalities due to its excellent diagnostic performance and rapid acquisition. However, it is also the largest contributor to radiation exposure from medical imaging [1]. For example, while a chest X-ray delivers an effective dose of approximately 0.1 mSv, a CT scan typically delivers between 1.5 and 15.4 mSv depending on the anatomical region, acquisition protocol, and number of phases [2]. In cases like multiphase or high-resolution protocols, radiation doses can reach up to 37 mSv, further emphasizing the need for dose reduction strategies.

Because the biological effects of ionizing radiation are cumulative and may damage genetic material, minimizing exposure while preserving diagnostic quality is a critical challenge. This concern is heightened for pediatric and young adult populations, who are more susceptible due to higher tissue radiosensitivity and a longer lifetime risk [1].

In this context, low-dose CT (LDCT) has emerged as a promising strategy to reduce radiation exposure while maintaining clinical utility. For instance, the National Lung Screening Trial reported a 20% reduction in lung cancer mortality among participants screened with low-dose CT (LDCT), using acquisition protocols with an average effective dose of approximately 1.5 mSv, compared to about 8 mSv in standard diagnostic CT [3]. Indeed, LDCT protocols aim to reduce radiation exposure by decreasing

[1] https://www.ncbi.nlm.nih.gov/books/NBK597564/

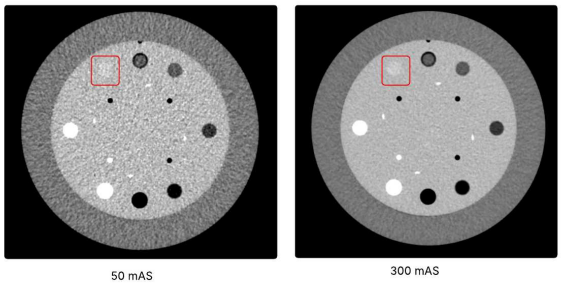


Fig. 1 Example of the effect of radiation dose reduction on a CT image using a CATPHAN-600 phantom. The image on the left shows a lower-dose CT scan with a higher noise level, making it difficult to detect small lesions.

X-ray intensity. However, this reduction also increases image noise, potentially compromising diagnostic reliability. Figure 1 illustrates the visual degradation caused by the reduction of dose in CT images acquired using a CATPHAN-600 phantom. This illustration evidences the difficulty of detecting small lesions in CT with a reduced dose. Therefore, the central challenge in LDCT imaging is to restore diagnostic image quality while preserving the benefits of dose reduction.

Traditional denoising and iterative reconstruction methods can partially alleviate this degradation, but their performance often deteriorates under extremely low-dose conditions. More recently, deep-learning-based methods have emerged as a powerful alternative for LDCT reconstruction, showing strong quantitative and visual performance [4–6]. In particular, generative approaches based on diffusion models have attracted significant attention due to their ability to model complex image distributions and progressively refine reconstructed images [7–9]. In parallel, Flow Matching (FM) has recently achieved strong performance across multiple generative modeling tasks, suggesting its potential as an alternative framework for medical image reconstruction.

Despite this progress, the current literature still presents important limitations. Existing LDCT reconstruction methods are often evaluated under inconsistent experimental settings, including different dataset splits, preprocessing pipelines, and noise generation procedures, which hinders fair comparison across approaches. In addition, many studies do not explicitly assess generalization across datasets with different acquisition characteristics, limiting the practical relevance of the reported results [5, 7, 9–14].

In this work, we address the aforementioned limitations by introducing a unified evaluation framework for LDCT reconstruction, making three contributions. First, we define standardized evaluation protocols, including the proposal of a new segmentation-based metric, to enable controlled comparisons across different reconstruction paradigms. Second, we present a benchmark between a discriminative model based on nnUnet, a GAN and a diffusion-based model. Finally, we evaluate the generalization capability of the reconstruction models across three heterogeneous public CT datasets, including both real low-dose acquisitions and synthetic low-dose images generated from full-dose CT scans.

To facilitate the reading, we organize this work as follows: Section 2 describes the related work. Section 3 presents the proposed methodology. Section 4 describes the experimental results. Finally, Section 5 describes the final remarks.

2 Related Work

Recent advances in artificial intelligence have driven the widespread adoption of deep learning in medical imaging, particularly in tasks such as segmentation, classification, and image reconstruction. In this context, deep learning methods can be broadly categorized into discriminative and generative models. Discriminative models learn mappings from input data to some specific output values (e.g., a set of labels), enabling tasks such as classification, regression or segmentation, whereas generative models aim to capture the underlying data distribution to synthesize realistic samples.

Among the various applications of deep learning in medical imaging, low-dose CT (LDCT) reconstruction has received significant attention due to its clinical relevance. This task can be formulated as an image-to-image translation problem, mapping low-dose scans to their normal-dose counterparts. Within this framework, discriminative models have been extensively explored, with U-Net-based architectures [15, 16] emerging as a standard baseline.

However, these discriminative-based approaches often struggle to generalize across datasets with varying conditions. Zhang et al. (2024) also found noticeable artifacts when training and evaluation resolutions differ, highlighting the limits of purely discriminative methods [17].

To overcome these limitations, researchers have explored generative models for LDCT reconstruction. These models enable flexible modeling of complex image distributions. Early works used generative adversarial networks (GANs); later approaches leverage diffusion and flow matching models.

2.1 Generative Adversarial Networks

Early GAN-based approaches focused on improving perceptual quality through adversarial training. Ma et al. [10] proposed a GAN-based model combining adversarial, structural similarity, and L1 losses, achieving structural similarity index (SSIM) and peak signal-to-noise ratio (PSNR) scores of 0.8454 and 28.49, respectively, though the evaluation was limited to a single patient.

Similarly, Kim et al. [11] employed a Wasserstein GAN (WGAN) with a VGG-based perceptual loss, reporting SSIM values below 0.80 on the LDCT Grand Challenge dataset; however, as with previous work, their evaluation remained very limited.

Building on these prior methodologies, Li et al. [12] introduced SACNN, incorporating attention mechanisms and a pretrained perceptual encoder within a WGAN framework. Despite these architectural improvements, evaluation was conducted on only two patients, limiting the statistical reliability of the reported performance.

More recently, as part of continuing efforts to enhance model robustness, Wang et al. [13] proposed an anatomy-aware GAN incorporating a semantic-guided contrastive learning strategy. While achieving improved results (PSNR = 33.71, SSIM = 0.9341),

the evaluation remains constrained to small datasets, and cross-dataset generalization is not thoroughly assessed.

2.2 Diffusion-based Models

Diffusion models have recently emerged as a powerful alternative for LDCT reconstruction, owing to their ability to model complex image distributions via iterative denoising. Xia et al. [5] applied Denoising Diffusion Probabilistic Models (DDPMs) to LDCT reconstruction and introduced an acceleration scheme based on modeling the reverse process with ordinary differential equations (ODEs), following the formulation of Lu et al. [18]. Their model, conditioned on the low-dose image, achieved substantial improvements in reconstruction quality while reducing inference time by up to 20×. They reported high results with a DDPM PSNR of 46.49 and an SSIM of 0.9873 across two patients.

In a complementary approach, Huang et al. [7] addressed the problem in the sinogram domain. They trained a diffusion model to infer the ideal sinogram from low-dose projections, using low-rank Hankel priors to regularize the reconstruction. This strategy allowed them to operate in the raw projection space before applying traditional image reconstruction techniques. They evaluated the proposal on the Mayo LDCT dataset, reporting PSNR and SSIM scores at three noise levels, achieving PSNR = 42.62 and SSIM = 0.9899 at the lowest noise level, and PSNR = 37.43 and SSIM = 0.9683 at the highest noise level.

Building on these advances, Gao and Shan [8] introduced CoCoDiff, a context-conditioned diffusion model that incorporates adjacent slices to preserve structural consistency. Their results show improved balance between noise suppression and CT value fidelity compared to GAN-based methods, which tend to introduce intensity shifts.

Extending this line of work, CoreDiff [9] further improved efficiency by reformulating the forward process to start from the low-dose image instead of pure Gaussian noise. This significantly reduces the number of required sampling steps while maintaining high reconstruction quality, achieving PSNR = 43.92 dB and SSIM = 0.9744 under a 10-step sampling regime.

Recently, Gong et al. [14] proposed a Flow Matching-based framework for ultra-low-dose CT denoising that addresses structural misalignment between low- and normal-dose images using an Image Purification (IP) strategy. This preprocessing step aligns anatomical structures, simplifying the learning problem and enabling training in the frequency domain. Quantitative results report PSNR = 31.85 dB and SSIM = 0.9179. Additionally, the IP strategy improves structural consistency and reduces reconstruction error, suggesting that aligning the input and target distributions facilitates a more stable and effective learning process. These results suggest that the primary gains arise from improved data alignment rather than the generative formulation itself.

Despite many proposed methods for low-dose CT reconstruction, several limitations hinder fair and consistent comparison. For example, many studies use only a small number of patients, which may introduce bias and reduce statistical reliability. Additionally, the lack of standardized evaluation protocols—including dataset splits

Method	Type	PSNR	SSIM	No. Patients
GAN (Ma et. al [10])	GAN	28.49	0.8454	1
WGAN (Kim et. al [11])	GAN	-	0.8000	1
SCANN (Li et. al [12])	GAN	22.18	0.7800	2
AA-GAN (Wang et. al [13])	GAN	33.71	0.9341	-
DDPM (Xia et. al [5])	Diffusion	46.49	0.9873	2
Sinogram+DDPM (Huang et. al [7])	Diffusion	42.62	0.9899	2
CoreDiff (Gao et. al [9])	Diffusion	43.92	0.9744	2
FM (Gong et. al [14])	Diffusion	31.85	0.9179	2

Table 1 Summary of the SOTA LDCT reconstruction models.

and preprocessing—makes direct comparison difficult. Further inconsistencies arise from differences in evaluation metrics, which range from intensity-based to perceptual and region-specific measures. To address these issues, this work aims to provide a unified evaluation framework for LDCT reconstruction. Specifically, it enables systematic comparison of discriminative and generative models, with a focus on GAN- and diffusion-based approaches. Furthermore, we highlight the need for controlled and reproducible experiments. To extend evaluation, we use noise simulation to analyze LDCT models on larger datasets, thus assessing model generalization.

3 Methodology

Our study focuses on three types of image-to-image translation models for low-dose CT reconstruction. We begin with a discriminative model using an nnUNet architecture. Next, we include a study of multichannel learning, in which CT slices are enriched with wavelet channels. Building on the best-performing discriminative model, we then examine the performance of generative adversarial networks. Finally, we turn our attention to the performance of conditional diffusion models.

To provide more reliable metrics, we define a noise simulation protocol and apply it to a public abdominal CT dataset, enabling us to assess the proposed reconstruction models on datasets distinct from those typically used for this task.

3.1 Problem Formulation

Low-dose CT (LDCT) reconstruction is formulated as an image-to-image translation problem, aiming to recover a normal-dose CT (NDCT) image from a low-dose input. Given a degraded observation x_{LD}, the objective is to estimate a reconstruction $\hat{x}_{ND}$ that preserves anatomical structures while reducing noise artifacts introduced by dose reduction.

3.2 Noise Simulation

To enable controlled evaluation and fair comparison, LDCT images were simulated from full-dose acquisitions. This approach is crucial both when paired data from low- and high-dose scans are unavailable and when evaluating generalization across datasets from different imaging protocols.

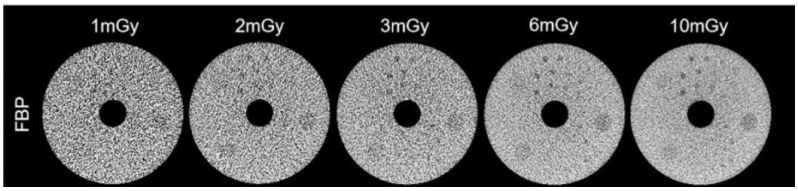


Fig. 2 CT images of a phantom acquired at different radiation dose levels. Lower mAs values lead to increased noise and degradation in image quality. [Adapted from [20]]

Furthermore, in CT imaging, radiation dose is closely related to the tube current–time product, commonly referred to as milliampere-seconds (mAs). Specifically, the mAs value directly controls the number of detected X-ray photons. As a result, lower mAs levels lead to fewer detected photons and increased noise in the reconstructed image, as illustrated in Fig. 2 [19]. This relationship forms the basis for simulating low-dose conditions.

Noise in CT mainly originates from the random (stochastic) nature of X-ray photon detection, mathematically described as a Poisson process (a model for counting random independent events). Additionally, modern CT systems introduce electronic noise, commonly represented as Gaussian noise (a model for random variations with a normal distribution). Based on these insights, and following Zeng et al. [21], we adopt the next two noise-simulation models:

- **Poisson noise**, which simulates the randomness of photon-limited acquisition.
- **Poisson-Gaussian noise**, which includes both photon (random event counts) and electronic (random electronic fluctuations) noise. In this case, the nousy signal is modeled as:

$$I = \text{Poisson}(\lambda) + \mathcal{N}(m_e, \sigma_e^2) \tag{1}$$

To further substantiate the connection between radiation dose and noise discussed above, real acquisitions using a Catphan 600 phantom were conducted across a range of tube current values (33–300 mAs). This experimental process validated the relationship between radiation dose and noise levels and complemented the noise modeling approach. The experimental analysis of the noise simulation is reported in Section 4.1.

3.3 Reconstruction Models

We evaluate three families of models for LDCT reconstruction: discriminative U-Net-based models, Generative Adversarial Networks (GANs), and diffusion-based generative models. Each approach is formulated as an image-to-image translation problem. The input is a low-dose CT (LDCT) image, and the target is its normal-dose counterpart.

3.3.1 UNet-based Reconstruction

We first explore two U-Net-based reconstruction models: one operating in the pixel domain and another in the wavelet domain. The pixel-domain model directly processes the LDCT image, learning a mapping to its normal-dose counterpart using a standard encoder–decoder U-Net architecture.

Alternatively, the wavelet-based model begins by applying a discrete wavelet transform (DWT), a mathematical technique that decomposes an image into components at different frequency levels. This process splits the input image into four frequency bands: LL (low-low), LH (low-high), HL (high-low), and HH (high-high). The LL component captures low-frequency structural information, while the LH, HL, and HH bands represent high-frequency details.

The network then takes as input the resulting multi-channel representation and learns to reconstruct the normal-dose image. This decomposition explicitly separates structural content from noise, simplifying the reconstruction task and improving the preservation of fine anatomical details. As a result, the wavelet-based formulation mitigates the over-smoothing effects commonly observed in pixel-domain reconstruction models. In addition, the pixel-domain model serves as a baseline, helping assess the impact of frequency-aware representations introduced by the wavelet-based approach.

3.3.2 Generative Adversarial Network

To improve perceptual quality and mitigate the over-smoothing behavior of purely discriminative models, we adopt a Generative Adversarial Network (GAN) framework operating in both the pixel and wavelet domains.

As in the U-Net baseline, the pixel-domain model directly processes the LDCT image to reconstruct its normal-dose counterpart. In the wavelet-based formulation, the input LDCT image is first decomposed using a discrete wavelet transform into four frequency bands (LL, LH, HL, HH). The generator, based on a U-Net architecture, takes this multi-channel representation as input and learns to reconstruct the corresponding normal-dose image.

The discriminator is designed to enforce structural consistency and perceptual realism. In particular, it operates using the low-frequency component (LL) as structural guidance, encouraging the generator to preserve anatomical integrity while reconstructing high-frequency details.

The training objective combines adversarial loss with reconstruction losses, including pixel-wise and structural similarity terms. This hybrid formulation allows the model to recover sharper details compared to standard U-Net approaches. However, GAN-based methods remain sensitive to training instability and may introduce artifacts or hallucinated structures, particularly under domain shifts. This dual formulation allows us to assess the impact of frequency-aware representations within adversarial training.

3.3.3 Diffusion-Based Models

We consider conditional diffusion models for LDCT reconstruction, where the input LDCT image guides the denoising process. The decision to use the LDCT as input for the conditioning is also motivated by the fact that an LDCT input conveys anatomical information that must be preserved through reconstruction. In addition, starting from an input with anatomical content could help reduce the number of reconstruction steps.

During our experiments, for all the evaluated diffusion-based models, we use a U-Net backbone with attention mechanisms, allowing it to capture both local structures and long-range dependencies.

To address the high computational cost typically associated with diffusion models, we evaluate multiple sampling strategies, including DDIM and high-order solvers such as DPMSolver, DPMSolver++, and UniPC. These methods significantly reduce the number of denoising steps required during inference while maintaining reconstruction quality. In addition, inspired by recent advances in this area, we include an optimized Flow Matching model called Residual Flow Matching (RFM) that directly transports LDCT images toward their paired target standard-dose images through a linear conditional path. As we show later, this strategy allows the model to achieve high performance with very few steps.

For the sake of fair comparisons, all models share the same conditioning input, enabling a consistent evaluation across architectures.

3.4 Evaluation Metrics

- **Peak Signal-to-Noise Ratio (PSNR)**: Peak Signal-to-Noise Ratio (PSNR) is determined by calculating the Mean Squared Error (MSE) between two images and then applying a logarithmic formula that compares this error to the maximum possible pixel value (MAX_I). PSNR is expressed in decibels (dB), with higher values indicating superior image reconstruction quality.

$$PSNR(a, b) = 10 \log_{10} \frac{MAX_I^2([a; b])}{MSE(a, b)}$$

- **Structural Similarity Index (SSIM)**: It assesses perceived image quality by comparing luminance, contrast, and structural information between a reference image and a distorted image. SSIM operates on local patches, typically using 11×11 Gaussian windows, to evaluate structural degradation instead of relying on pixel-by-pixel differences. The resulting SSIM value ranges from -1 to 1, with values closer to 1 indicating higher similarity.

$$SSIM(a, b) = \frac{(2\mu_a\mu_b + C_1)(2\sigma_{ab}^2 + C_2)}{(\mu_a^2 + \mu_b^2 + C_1)(\sigma_a^2 + \sigma_b^2 + C_2)}$$

  where σ_{ab} is the cross-variance and C_1, C_2 are stabilization constants that we set to.
- **Throughput**: To evaluate computational efficiency, we measure the throughput of each model in terms of the number of reconstructed images per second. This metric is particularly relevant for clinical applications, where reconstruction time is a critical constraint. Throughput is defined as:

$$Throughput = \frac{N}{T}$$

where N is the number of processed images and T is the total inference time in seconds. All measurements are performed under the same hardware conditions to ensure a fair comparison across models.

- **Segmentation Score**: Using TotalSegmentator [22] as a segmentation model, we evaluate how well it segments the liver in abdominal CTs from regular-dose scans and from reconstructed low-dose scans.

3.5 Statistical Analyses

Statistical analyses and graphical representations were performed using GraphPad Prism 6 (GraphPad Software Inc., La Jolla, CA, USA). Comparisons among three or more groups were conducted using one-way analysis of variance (ANOVA), followed by Tukey's post hoc test when appropriate. Comparisons between two groups were performed using Student's t-test. In addition, simple linear regression analysis was performed to evaluate the agreement between the simulated noise and real low-dose images. A p-value < 0.05 was considered statistically significant.

3.6 Datasets

In all our experiments, we use the following three publicly available datasets.

- TCIA LDCT and Projection data: This is a dataset with CT images of 298 patients, including liver, chest, and head scans. Only a portion of all patients have full-dose images and their corresponding quarter-dose images. Taking advantage of the full-dose and quarter-dose pairs, we use 120 patients for training (45 chest, 36 head and 39 liver) and 30 for validation (5 chest, 14 head and 11 liver). The whole dataset can be downloaded from https://www.cancerimagingarchive.net/collection/ldct-and-projection-data/.
- Mayo Grand Challenge: This is dataset with liver CT images from 30 patients, with 10 patients for training and 20 patients for test, with only the training patients having full dose images and their corresponding quarter dose images. The complete dataset can be downloaded from https://aapm.app.box.com/s/eaw4jddb53keg1bptavvvd1sf4x3pe9h/folder/144594475090.
- ColoRectal Liver Metastases (CRLM): A dataset of CT images from 197 patients; the original data only include full-dose images, so artificial noise will be added to obtain quarter-dose images, as well as half- and one-eighth-dose images. The whole dataset can be downloaded from https://www.cancerimagingarchive.net/collection/colorectal-liver-metastases.

4 Results and Discussion

4.1 Noise Synthesis

The results obtained from the proposed noise synthesis and reconstruction approaches are presented and discussed in this section.

Figure 3 presents a qualitative comparison between simulated and real low-dose CT images at two dose levels. Figure A shows the comparison between the high-dose

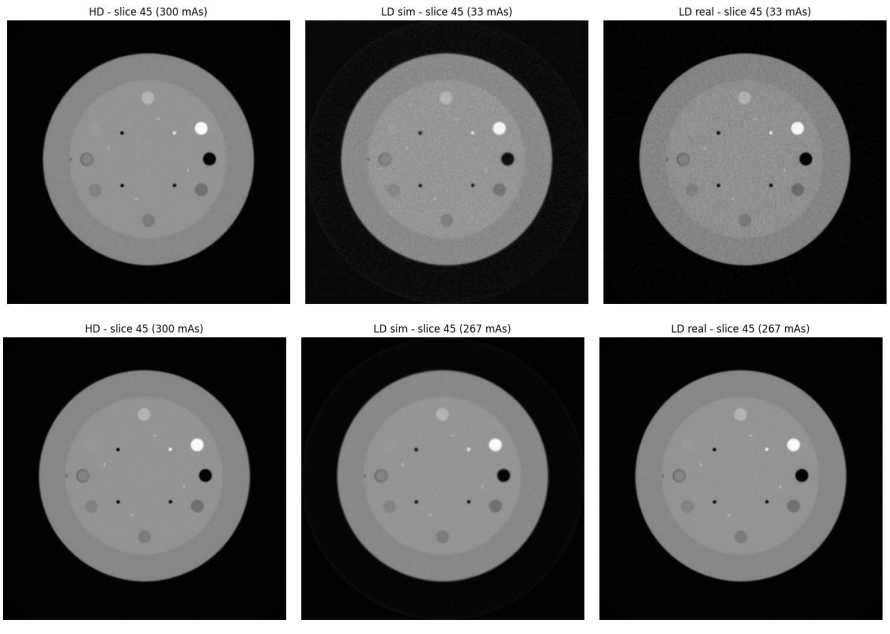


Fig. 3 Visual comparison between high-dose (300 mAs), simulated low-dose, and real low-dose CT images for (A) 33 mAs and (B) 267 mAs acquisitions.

(HD) reference image acquired at 300 mAs, the simulated low-dose image, and the real low-dose acquisition at 33 mAs. At this lower dose level, a pronounced increase in noise is observed in both simulated and real images relative to the HD reference. Despite the higher noise magnitude, the simulation reproduces the overall noise texture and contrast distribution seen in the real acquisition, maintaining visibility of the main phantom structures and inserts.

Figure B shows the same comparison for the 267 mAs dose level. In this case, both simulated and real low-dose images exhibit reduced noise compared to Figure A, appearing closer to the HD reference. The simulated image demonstrates strong visual agreement with the real low-dose acquisition, particularly in terms of background uniformity and contrast preservation. Overall, the results indicate that the proposed noise synthesis method realistically reproduces dose-dependent noise characteristics across different mAs levels, from very low to near-high-dose conditions.

Using all images acquired at different mAs values, the SNR and CNR were calculated. Figure 4 summarizes the comparison between real CT images and simulated low-dose images generated using the two proposed noise synthesis strategies. Scatter plots (A–D) present the relationship between SNR and CNR values measured from real images versus simulated images, while plots (E–F) show the differences between methods.

Panels (A) and (B) present the SNR comparison for Method 01 (Poisson–Gaussian noise) and Method 02 (Poisson-only noise), respectively. In both cases, a strong linear relationship is observed between real and simulated measurements, with coefficients of determination ($R^2 = 0.9957$) and ($R^2 = 0.9956$), indicating excellent agreement.

The regression slopes ((∼ 1.074)) suggest that both simulation methods slightly overestimate SNR values compared to real acquisitions.

Similarly, panels (C) and (D) show the CNR comparison for Methods 01 and 02. High correlations were also obtained (($R^2 = 0.9952$) and ($R^2 = 0.9951$)), demonstrating that both approaches accurately reproduce contrast-related image characteristics. However, the regression slopes (~ 0.76) indicate a systematic underestimation of CNR in the simulated data relative to real measurements.

The difference plots shown in panels (E) and (F) further compare the performance of both simulation approaches. For both SNR and CNR, no statistically significant differences (ns) were observed between Method 01 and Method 02, suggesting that excluding the Gaussian component does not significantly alter the overall quantitative image quality metrics under the conditions evaluated. Statistical analysis was performed using a paired Student's t-test.

Overall, these results indicate that both noise synthesis strategies are capable of reproducing the statistical behavior of real low-dose CT images, with comparable performance in terms of SNR and CNR metrics.

4.2 Multichannel Training

Leveraging the high throughput demonstrated by discriminative models, we used a U-Net-based model to determine whether frequency-domain information from the Wavelet Transform (WT) could improve performance compared to using only low-dose CT (LD). To this end, we evaluate the U-UNet model's performance on LD images and when we incorporate WT, by concatenation. Our results show that both strategies achieve similar PSNR and SSIM values. The results using LD-only are 42.87 and 0.9716 for both metrics, while with LD+WT are 42.96 and 0.9734. Therefore, since adding more information at the input just makes the model more complex, we use the LD images as input hereafter.

4.3 Performance on the TCIA's Validation Dataset

Table 2 shows the performance of the evaluated models on the TCIA's validation dataset. Here, the diffusion-based models outperform GANs and UNets in PSNR and SSIM metrics, for a number of reconstruction steps greater than or equal to 10. However, we observe more efficient performance of the RFM model, which can achieve good results even with one step only. Indeed, RFM-1 gets 50.98 and 0.9865 for PSNR and SSIM, respectively, very similar results to DDPM-10 or DDIM-10. The behavior of RFM for one step is highly remarkable considering that DDPM and DDIM cannot reconstruct LDCT for one step.

Furthermore, as we detail later, we have run statistical tests to determine whether diffusion models provide a significant improvement over other models. Our results, summarized in Figure 6, show that for the TCIA dataset RFM-1, DDPM-10 and DDIM-10 produce significant improvement over UNet-LD or GAN models in terms of PSNR, but when comparing structural similarity, the performance of diffusion-based models and UNet is not significantly different.

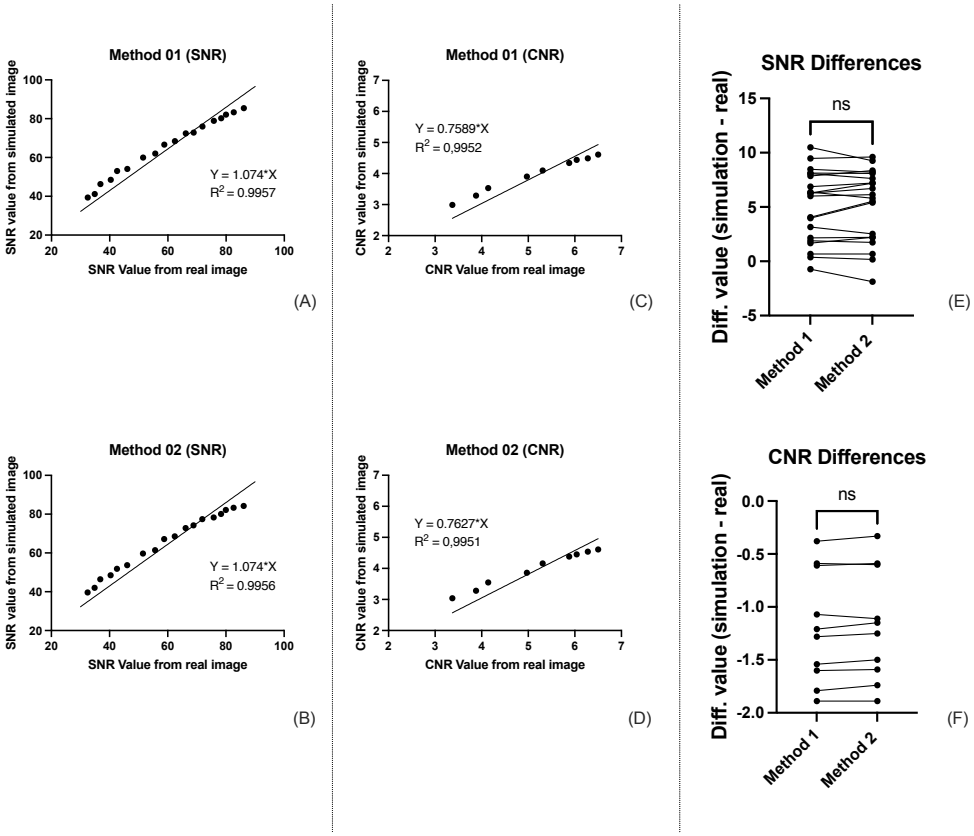


Fig. 4 Correlation analysis between real and simulated images for SNR and CNR using Method 01 and Method 02 (A–D), with paired differences shown in (E–F).

At this point, it is important to remember what each metric measures. PSNR focuses on the pixel-wise difference between the ground-truth and the inferred versions, while SSIM focuses on the consistency of structural information between the two versions of an image. Beyond the exact pixel values, our study aims to evaluate the consistency of the reconstructed CT's structural components, prioritizing SSIM values. Although we found that UNet and diffusion models perform similarly in terms in structural similarity, it is critical to know if this behavior still holds for other datasets (inter-dataset generalization). Later, we will focus on this kind of analysis.

However, although DDPM, DDIM and UNet achieve similar structural information with SSIM values around 0.98, there is a critical aspect we need to take into account: the models' efficiency. The iterative nature of diffusion-based models makes them slower than one-run models like UNet or GAN. Our experiments show that diffusion models with a low number of iterations are still two orders of magnitude slower, processing at 6 FPS, compared to the 700 FPS of UNet or GAN.

Model	**PSNR**	**SSIM**	Throughput (fps)
UNet LD	42.8673 ± 7.1037	0.9716 ± 0.0296	736.97
UNet LD+WT	42.9594 ± 6.9225	0.9734 ± 0.0287	720.71
GAN LD	39.1944 ± 4.8664	0.9443 ± 0.0282	719.04
GAN LD+WT	37.4163 ± 3.9494	0.9582 ± 0.0280	719.04
DDPM 1	6.7882 ± 0.1868	0.0034 ± 0.0009	95.55
DDPM 10	50.9196 ± 7.5389	0.9847 ± 0.0176	9.26
DDPM 25	50.8155 ± 7.4242	0.9846 ± 0.0178	3.72
DDPM 150	50.7225 ± 7.6807	0.9837 ± 0.0189	0.62
DDPM 500	50.7198 ± 7.6783	0.9837 ± 0.0189	0.19
DDPM 1000	50.7183 ± 7.6779	0.9837 ± 0.0189	0.093
DDIM 1	6.7882 ± 0.1868	0.0034 ± 0.0009	96.01
DDIM 10	51.1688 ± 7.3179	0.9860 ± 0.0158	8.65
DDIM 25	50.8155 ± 7.4242	0.9846 ± 0.0177	4.07
DDIM 150	50.0121 ± 6.9641	0.9830 ± 0.0193	0.66
DDIM 500	49.6813 ± 6.7893	0.9825 ± 0.0197	0.19
DDIM 1000	48.8896 ± 6.3559	0.9811 ± 0.0198	0.09
RFM 1	50.9800 ± 7.2000	0.9865 ± 0.0149	94.54
RFM 4	50.6700 ± 7.3200	0.9851 ± 0.0163	23.32
RFM 10	50.2700 ± 7.4900	0.9829 ± 0.0191	9.31

Table 2 Results on validation dataset TCIA (N=30 patients) including UNet-based models, GAN and diffusion.

Leveraging CTs from different body parts in TCIA, we also explore the performance of reconstruction models for three different body regions: liver, chest and brain (neuro). The results of this study are presented in Table 3, with PSNR and SSIM values reported. In this analysis, the chest is the most difficult part, where all the models achieve their worst performance in both studied metrics.

The marked variation in reconstruction performance among the three body parts prompted us to investigate the underlying causes. Upon examining the CT scans for each group, we identified a substantial difference in noise levels. Specifically, the average signal-to-noise ratio (SNR) score computed for each group revealed a pronounced decrease for chest CTs, with SNR = 22.47, compared to liver CTs (SNR = 42.75) and brain CTs, which achieved the highest value (SNR = 54.42). Figure 5 presents the noise levels for each anatomical region studied.

4.4 Inter-Dataset Performance

Generalization across datasets is becoming a critical property of a machine learning model. Indeed, it is necessary to consider models that perform well on datasets with distributions that differ from those of the training set to obtain a more accurate estimate of performance in new environments. Accordingly, our evaluation protocol is extended to incorporate two additional testing datasets, Mayo and CRLM, as explained in Section 3.6.

Table 4 presents the performance of the evaluated models on the Mayo dataset, which demonstrates a performance pattern similar to that observed in TCIA. Here, diffusion-based models maintain superior performance, with a more pronounced difference in terms of PSNR. And, again, RFM-1 reaches comparable performance to

	Liver		Chest		Brain	
Model	**PSNR**	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**
UNet LD	47.3014 ± 2.2580	0.9937 ± 0.0020	35.3807 ± 1.7669	0.9414 ± 0.0200	53.5157 ± 2.1026	0.9989 ± 0.0004
UNet LD+WT	47.3917 ± 2.0896	0.9951 ± 0.0014	35.6308 ± 1.8533	0.9442 ± 0.0202	53.1724 ± 1.7159	0.9981 ± 0.0005
GAN LD	43.0009 ± 1.7300	0.9661 ± 0.0093	33.9552 ± 1.5501	0.9199 ± 0.0240	43.9165 ± 1.5813	0.9505 ± 0.0168
GAN LD+WT	40.7281 ± 1.2450	0.9795 ± 0.0068	33.2459 ± 1.7239	0.9320 ± 0.0230	40.1212 ± 1.559	0.9731 ± 0.0092
DDPM 10	54.9467 ± 1.6594	0.9976 ± 0.0009	43.2048 ± 1.6076	0.9673 ± 0.0135	63.9944 ± 3.6188	0.9996 ± 0.0003
DDPM 25	54.8995 ± 1.6599	0.9975 ± 0.0009	43.1593 ± 1.6207	0.9671 ± 0.0137	63.452 ± 3.3160	0.9996 ± 0.0003
DDPM 150	54.8770 ± 1.7339	0.9975 ± 0.0009	42.8398 ± 1.6531	0.9651 ± 0.0148	63.9279 ± 3.5925	0.9996 ± 0.0003
DDPM 500	54.8750 ± 1.7343	0.9975 ± 0.0009	42.8373 ± 1.6551	0.9651 ± 0.0148	63.9219 ± 3.5531	0.9996 ± 0.0003
DDPM 1000	54.8753 ± 1.7347	0.9975 ± 0.0009	42.8356 ± 1.6547	0.9651 ± 0.0148	63.9146 ± 3.5595	0.9996 ± 0.0003
DDIM 10	55.0618 ± 1.5872	0.9976 ± 0.0008	43.6867 ± 1.5645	0.9703 ± 0.0119	63.8992 ± 3.5291	0.9996 ± 0.0003
DDIM 25	54.8995 ± 1.6599	0.9975 ± 0.0009	43.1593 ± 1.6207	0.9671 ± 0.0137	63.4523 ± 3.3160	0.9996 ± 0.0003
DDIM 150	54.2185 ± 1.7793	0.9970 ± 0.0012	42.7383 ± 1.6687	0.9642 ± 0.0153	60.7513 ± 3.6835	0.9990 ± 0.0011
DDIM 500	53.8388 ± 1.9512	0.9966 ± 0.0016	42.6349 ± 1.6821	0.9634 ± 0.0157	59.7639 ± 4.1986	0.9986 ± 0.0017
DDIM 1000	52.8238 ± 2.6060	0.9952 ± 0.0040	42.5402 ± 1.6642	0.9623 ± 0.0159	56.8498 ± 5.2221	0.9967 ± 0.0040
RFM Base 1	54.7394 ± 1.8218	0.9975 ± 0.0009	43.6774 ± 1.4895	0.9715 ± 0.0107	63.5297 ± 3.6519	0.9996 ± 0.0003
RFM Base 4	54.4395 ± 1.8656	0.9973 ± 0.0010	43.2858 ± 1.4905	0.9686 ± 0.0115	63.5370 ± 3.9246	0.9996 ± 0.0004

Table 3 Results of the reconstruction models on TCIA validation dataset for three different body parts: liver (N_{liver}=11), chest (N_{chest}=5), and brain (N_{brain}=14)

.

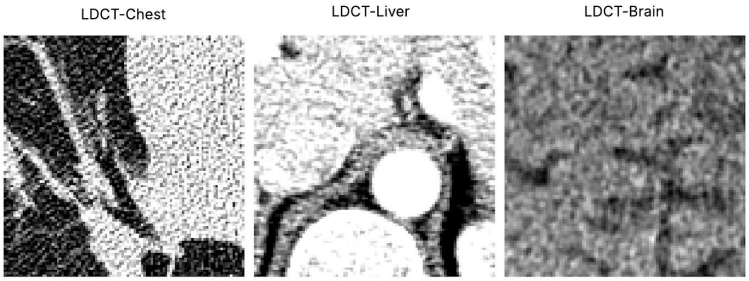


Fig. 5 A sample of LDCT for each body part. From left to right, we have chest (SNR = 23.80), liver (SNR = 42.78), and brain (54.43).

DDIM-10 or DDPM-10. RFM-1 obtains a value of 48.83 for PSNR and 0.9866 for SSIM.

On the other hand, Table 5 presents the reconstruction results for the CRLM (abdominal scans) dataset, showing a broader difference in favor of using diffusion models. Here, UNet-based and GAN-based models perform very similarly, with p-values greater than 0.6 between them. However, both present a broader difference from the diffusion-based models. In particular, DDIM-10 outperforms UNet by 11 points in PSNR, achieving a score of 39.40, and by 6% in SSIM, achieving a score of 0.9464, when the dose reduction is 0.5. Here, we note the superiority of DDIM over RFM, where the latter is below the former, even for the same number of reconstruction steps.

Taking advantage of the noise simulation described in Section 4.1, we explore the capacity of the reconstruction models for three different reduction doses (1/2, 1/4, 1/8). We observe very small variations in the results for the studied reduced doses, with diffusion models remaining the best methods, with SSIM values over 0.98 PSNR around 40 for DDIM-10, even at 1/8 of the regular doses.

Model	PSNR	SSIM
UNet LD	42,2123 ± 3.1468	0.9814 ± 0.0107
UNet LD+WT	41,1185 ± 3.1928	0.9782 ± 0.0130
GAN LD	40.2121 ± 2.5386	0.9693 ± 0.0124
GAN LD+WT	38.4888 ± 1.9146	0.9723 ± 0.0118
DDPM 10	48.9561 ± 2.7778	0.9869 ± 0.0075
DDPM 25	48.8271 ± 2.8373	0.9861 ± 0.0082
DDPM 150	48.7706 ± 2.8602	0.9858 ± 0.0085
DDPM 500	48.7678 ± 2.8591	0.9858 ± 0.0085
DDPM 1000	48.7681 ± 2.8644	0.9857 ± 0.0085
DDIM 10	49.2139 ± 2.6839	0.9881 ± 0.0066
DDIM 25	48.9549 ± 2.7615	0.9869 ± 0.0075
DDIM 150	48.5486 ± 2.7508	0.9854 ± 0.0085
DDIM 500	48.3369 ± 2.7612	0.9846 ± 0.0089
DDIM 1000	47.8954 ± 2.7453	0.9831 ± 0.0098
RFM 1	48.8278 ± 2.8393	0.9866 ± 0.0073
RFM 4	48.5768 ± 2.7841	0.9853 ± 0.0081
RFM 10	48.3408 ± 2.7506	0.9847 ± 0.0083

Table 4 Results on the MAYO validation dataset($N = 10$). Here, diffusion-based models beast UNets and GANs.

	1/2		1/4		1/8	
Model	**PSNR**	**SSIM**	**PSNR**	**SSIM**	**PSNR**	**SSIM**
UNet LD	28,5707 ± 4,4288	0,8893 ± 0,0664	29,0268 ± 4,4727	0,8938 ± 0,0669	29,5138 ± 4,4448	0,8940 ± 0,0698
UNet LD+WT	28,1575 ± 4,4224	0,8656 ± 0,0749	28,4060 ± 4,4739	0,8644 ± 0,0777	28,6399 ± 4,4952	0,8607 ± 0,0823
GAN LD	28,2120 ± 4,3780	0,8879 ± 0,0649	28,7339 ± 4,4259	0,9013 ± 0,0619	29,2802 ± 4,3840	0,9112 ± 0,0600
GAN LD+WT	28,8034 ± 4,7197	0,8715 ± 0,0780	28,9621 ± 4,7833	0,8703 ± 0,0822	29,0740 ± 4,8292	0,8674 ± 0,0883
DDPM 10	37.3937 ± 2.8355	0.9318 ± 0.0368	37.9696 ± 2.9122	0.9404 ± 0.0377	38.5301 ± 2.8451	0.9457 ± 0.0402
DDPM 25	35.3986 ± 3.7807	0.9061 ± 0.0505	36.1279 ± 3.8794	0.9191 ± 0.0496	36.9123 ± 3.757	0.9303 ± 0.0507
DDPM 150	34.3673 ± 4.6479	0.8879 ± 0.0651	35.1037 ± 4.7697	0.9014 ± 0.0631	35.9568 ± 4.6362	0.9157 ± 0.061
DDPM 500	34.3325 ± 4.6837	0.8872 ± 0.0658	35.0550 ± 4.8310	0.9005 ± 0.064	35.8976 ± 4.7411	0.9145 ± 0.0631
DDPM 1000	34.3310 ± 4.6838	0.8872 ± 0.0658	35.0557 ± 4.8329	0.9005 ± 0.0642	35.8947 ± 4.7447	0.9144 ± 0.0632
DDIM 10	39.4040 ± 2.2475	0.9464 ± 0.0313	39.8304 ± 2.2929	0.9518 ± 0.0325	40.2076 ± 2.2674	0.9544 ± 0.0346
DDIM 25	37.0598 ± 3.1159	0.9223 ± 0.0470	37.6804 ± 3.1844	0.9324 ± 0.0477	38.2991 ± 3.0971	0.9395 ± 0.0504
DDIM 150	34.7708 ± 4.3282	0.8916 ± 0.0654	35.5088 ± 4.4123	0.9052 ± 0.0633	36.3280 ± 4.2744	0.9181 ± 0.0628
DDIM 500	34.3774 ± 4.6212	0.8864 ± 0.0660	35.0938 ± 4.7460	0.8993 ± 0.0647	35.9102 ± 4.6434	0.9126 ± 0.0639
DDIM 1000	34.2837 ± 4.6264	0.8834 ± 0.0664	34.9876 ± 4.7612	0.8963 ± 0.0656	35.7956 ± 4.6800	0.9097 ± 0.0647
RFM 1	33.8684 ± 4.7908	0.8778 ± 0.0736	34.3378 ± 4.9529	0.8838 ± 0.0753	34.9610 ± 5.0536	0.8921 ± 0.0780
RFM 4	33.8602 ± 4.7414	0.8771 ± 0.0734	34.2663 ± 4.8741	0.8811 ± 0.0749	34.7736 ± 4.9556	0.8862 ± 0.0776
RFM 10	33.8500 ± 4.7241	0.8768 ± 0.0734	34.2327 ± 4.8416	0.8802 ± 0.0750	34.7016 ± 4.9093	0.8842 ± 0.0778

Table 5 Results on the CRLM dataset (N=96). Here, we show the performance of the reconstruction models for three percentages of reduction doses (1/2, 1/4, and 1/8). In general, we observe very small variations in the results for the studied doses.

4.5 Statistical Analysis

Figure 6 summarizes the statistical comparison of the evaluated reconstruction models in terms of PSNR and SSIM across the three datasets considered in this study (TCIA, Mayo, and CRLM). DDPM-10 achieved PSNR and SSIM values comparable to those of DDIM-10, with no statistically significant differences between both methods. Based on these findings, DDPM-10 was omitted from the subsequent figure.

For the TCIA validation dataset, DDIM-10 and RFM-1 achieved significantly higher PSNR values than both UNet and GAN approaches ($p < 0.0001$), reaching approximately 51 dB compared with 43 dB for UNet and 39 dB for GAN. However,

despite these substantial gains in pixel-wise fidelity, the differences in SSIM were considerably smaller. All models achieved high structural similarity values ($SSIM > 0.94$), and no statistically significant differences were observed between DDIM-10 and RFM-1 or between DDIM-10 and UNet, suggesting that all methods were capable of preserving the main anatomical structures present in the images. Additionally, GAN showed significantly lower SSIM values than both UNet and DDIM-10/RFM-1, which achieved SSIM values of approximately 0.985.

A similar trend was observed in the Mayo dataset. DDIM-10 and RFM-1 again produced significantly higher PSNR values, reaching approximately 49 dB, compared with 42 dB and 40 dB for UNet and GAN, respectively. Nevertheless, SSIM values remained very close among methods, ranging from approximately 0.97 to 0.99. No statistically significant differences were observed between DDIM-10 and RFM-1 in either PSNR or SSIM. Although some statistical differences were detected among the remaining models, their magnitude was relatively small compared with the PSNR improvements. These findings indicate that DDIM-10 and RFM-1 primarily improve pixel-level accuracy while providing only modest gains in structural similarity when evaluated on datasets that are relatively similar to the training distribution.

The most pronounced differences were observed in the CRLM dataset, which represents an out-of-distribution evaluation scenario. Initially, synthetic low-dose images were generated using dose reductions of 1/2, 1/4, and 1/8; however, only minor variations were observed across these dose levels, indicating that the relative ranking of the reconstruction models remained stable even under more aggressive dose reductions. Therefore, subsequent statistical analyses were focused on the 1/4-dose condition (similar to the other datasets). In this dataset, DDIM-10 demonstrated substantially better performance than both UNet and GAN. In particular, DDIM-10 achieved a PSNR of approximately 39.8 dB, about 10 dB higher than UNet and GAN, while also reaching an SSIM of approximately 0.95. Although RFM-1 achieved PSNR values comparable to those of UNet and GAN, its SSIM remained significantly lower than that of DDIM-10, indicating reduced structural preservation under domain-shift conditions. Unlike the TCIA and Mayo datasets, significant differences were observed not only in PSNR but also in SSIM, indicating that DDIM-10 preserved anatomical structures more effectively under out-of-distribution conditions. These results suggest that the advantages of diffusion-based reconstruction become more evident as the evaluation moves further away from the training distribution, highlighting their superior robustness and generalization capability across heterogeneous datasets.

4.6 A Semantic Metric

PSNR and SSIM, although traditional metrics for assessing image quality, are rarely good measures in imagenology, where semantic content is the critical issue. A better performance measurement would be the evalution made by radiologists, but it is impractical. Even though we can access some radiologists, they could not evaluate a large amount of images; we would need an automatic tool for that. Fortunately, we now have access to high-performance computer vision models for semantic tasks on CTs. Therefore, we propose using an automatic segmentation model to assess the quality of the reconstructed CTs. To validate that this metric provides valuable information,

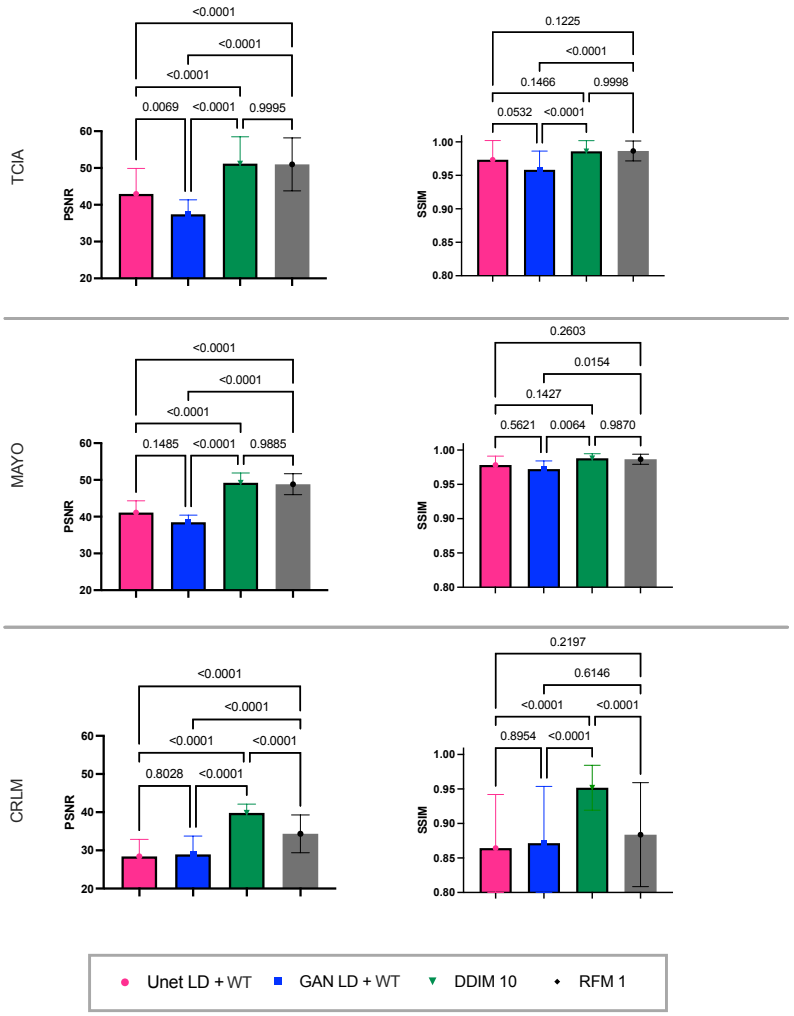


Fig. 6 Statistical analysis on the reconstruction models in terms of PSNR (left column) and SSIM (right column) for the different evaluated datasets. [1st row: TCIA, 2nd row: Mayo, 3rd: CRLM]. Horizontal brackets indicate pairwise comparisons between groups, and the associated p-values obtained from Tukey's post hoc test are shown above each bracket. Note that when we evaluated on an out-of-distribution dataset, diffusion models show a clearer performance difference compared to UNet and GAN.

we first evaluate the capacity of the selected segmentator to segment different parts of an abdominal CTs (livers, vessels and tumors) on full-dose and low-dose scans. The segmentator achieves 94% DICE for liver segmentation on full-dose scans, but its performance drops to 82% on low-dose scans. For the semantic segmentator, we can choose any SOTA model. In this work, we use Total Segmentator [22], which has shown outstanding performance across different body regions.

Table 6 show the results of the segmentation-based metric for the diverse reconstruction methods. We observe that DDIM outperforms the others, achieving a Dice score of 0.9099, surpassing UNet or GAN models by 5%. Furthermore, RFM, even though it is efficient, has lower effectiveness than DDIM.

Model	Liver (Dice)
Full Dose	0.9447 ± 0.0317
Low Dose	0.8249 ± 0.1666
UNet LD	0.8556 ± 0.1241
UNet LD+WT	0.8519 ± 0.1279
GAN LD	0.8351 ± 0.1352
GAN LD+WT	0.8508 ± 0.1239
DDPM 25	0.8878 ± 0.0907
DDIM 25	0.9099 ± 0.0693
RFM 1	0.8559 ± 0.1197
RFM 4	0.8534 ± 0.1214
RFM 10	0.8519 ± 0.1238

Table 6 Liver segmentation performance on the reconstructed CTs for the CRLM dataset. Here, we highlight the effectiveness of DDIM-25, which achieves a Dice of 0.9099, which is the closest to the 0.9447, achieved when using full-dose CTs.

4.7 Qualitative Results by Radiologists

Figure 7 illustrate a sample of reconstructed CTs produced by DDDIM-25 highlighting the reconstruction of small structures. Moreover, a qualitative evaluation of the dataset used for the segmentation tests was performed by two radiologists with over 10 years of experience. In general, the reconstructed images showed adequate anatomical consistency across slices within the same patient. However, a smoothed appearance of the images was observed, with some loss of contour definition, a finding expected in noise-reduction techniques.

Furthermore, variability in the representation of focal liver lesions was identified in the CRLM dataset. In particular, in some cases of liver metastases with vascularized components, the reconstructed images showed a decrease or partial loss of the areas of greatest attenuation. While this phenomenon did not appear to significantly compromise the detectability of the lesions, it could affect radiological interpretation of certain findings, particularly in characterizing their internal components and in differential diagnosis.

Figure 8 describes three representative cases (1193, 1046 and 1128) from the CRLM dataset. Each columns shows the original images, the the Low-Dose CT images, and the image reconstructed using DDIM-25, respectively. Thus, the first row shows a hypovascular metastasis in the left hepatic lobe, which is adequately represented in

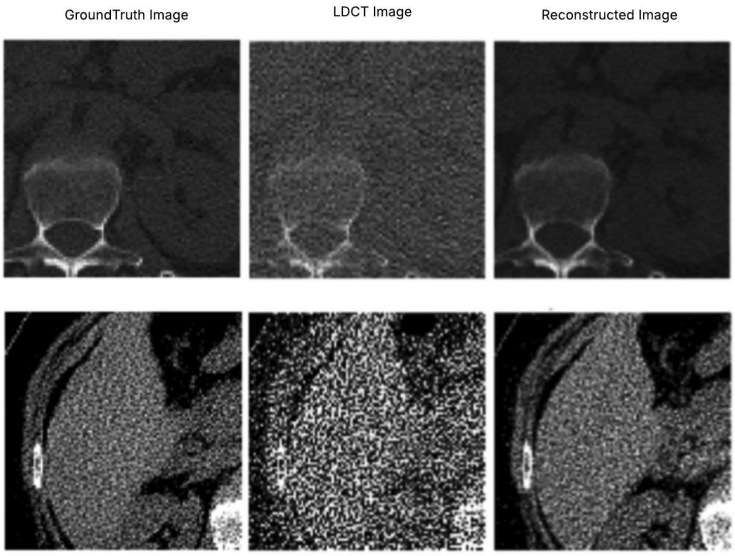


Fig. 7 Qualitative results: The second and fourth rows got closer to local structures, where the diffusion-based model achieves a structural similarity index measure (SSIM) of 0.8946 and a peak signal-to-noise ratio (PSNR) of 32.73 dB. For the zoomed-in images, we use a windowing strategy with W=200 and L=50.

the reconstructed image. The second row shows a patient with a metastasis in the hepatic dome, with a peripheral vascularized component. Note that, in the reconstructed image, the lesion remains faintly visible, although with partial loss of the peripheral vascularized structure. Finally, the third row shows a lesion in the left hepatic lobe in a liver with steatosis, a condition that alters the lesion's radiological appearance. The lesion appears relatively isodense with respect to the hepatic parenchyma and maintains adequate representation in the reconstructed image.

4.8 Challenges

After analyzing the experimental results, we can infer the following two main challenges:

- **Semantic Effectiveness**: We observe that the diffusion models achieve 88% DICE in liver segmentation, outperforming GAN and UNet but 4 points below the full-dose results. Therefore, there is still a gap to be filled by better models. Moreover, it is critical to extend the evaluation to analyze finer structures, such as small tumors or vessels.
- **Throughput**: Because of their iterative nature, diffusion models are critically slow. Thus, we need to improve efficiency by leveraging modern ideas in image generation, such as latent-space techniques.

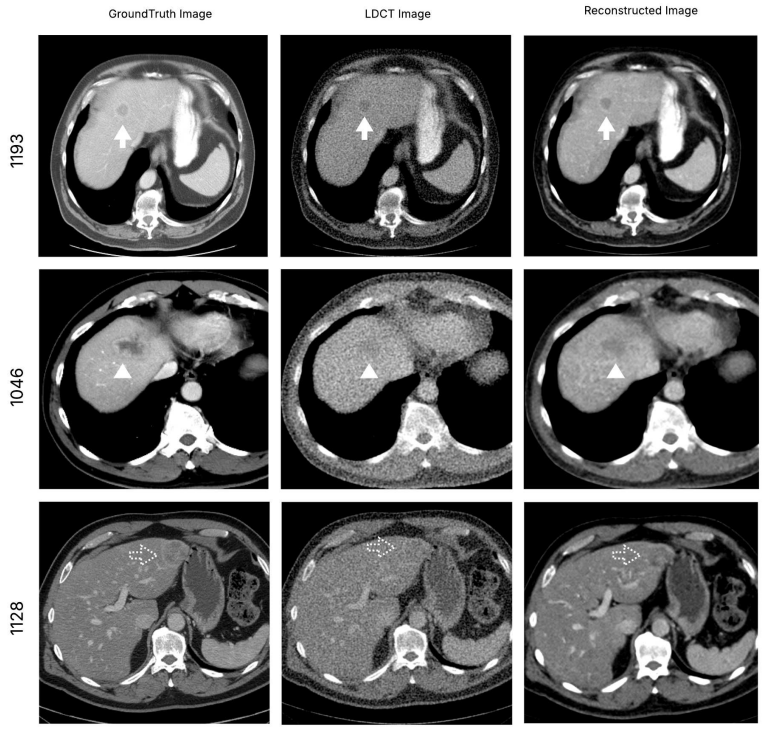


Fig. 8 Case of Study: Each case (1193, 1046, 1128) is presented by row, reporting the ground-truth image, the low-dose image and the reconstructed image.

5 Conclusions

Given advances in artificial intelligence and computer vision, and recognizing the importance of reducing radiation dose in CT scans, we have presented a comprehensive fair study of modern generative models for LDCT reconstructions. Beyond effectiveness on a single dataset, our main goal is to assess the models' generalization across datasets. Moreover, we propose a new metric based on semantic segmentation, in addition to traditional image quality metrics.

Our results show that diffusion-based models with a low number of iterations outperform UNet and GAN, especially when tested on out-of-distribution datasets. We also observe that a DDIM-10 is sufficient to outperform UNet and GAN by 10 points in PSNR and 6 points in structural similarity (SSIM). Furthermore, it is important to highlight that DDIM-10 resulted in 4 points below the full-dose images when CTs were exposed to an automatic liver segmentation process. In contrast, UNet and GAN dropped by up to 9 points below the full-dose.

We also introduce RFM, which, although it accelerates reconstruction, suffers from low inter-dataset generalization, with performance similar to UNet or GAN.

However, despite their high reconstruction fidelity and better generalization, diffusion models remain computationally demanding due to their iterative nature, motivating the exploration of more efficient formulations, such as latent generative models.

Declarations

Author Contribution

JMS conceptualized, proposed the methodology, designed the experiments, wrote the first draft and supervised the research. TS implemented the diffusion models, evaluated and improved the writing. VC conducted the formal validation and improved the writing. AX evaluated the noise simulation method and ran statistical tests. HH and YO conducted clinical validation and reviewed the manuscript. JD validated the results and reviewed the manuscript. CF, JC and JS implemented the evaluated models and ran evaluation experiments.

Research Funding

This research was supported by the National Agency of Research and Development, Chile, under the grants 1260784 and 13250156.

Data Availability

The datasets used for this work are publicly available at:

- **TCIA** https://www.cancerimagingarchive.net/collection/ldct-and-projection-data/
- **Mayo Grand Challenge** https://aapm.app.box.com/s/eaw4jddb53keg1bptavvvd1sf4x3pe9h/folder/144594475090
- **CRLM** https://www.cancerimagingarchive.net/collection/colorectal-liver-metastases

References

[1] Brenner, D.J., Hall, E.J.: Computed Tomography — An Increasing Source of Radiation Exposure. New England Journal of Medicine **357**(22), 2277–2284 (2007) https://doi.org/10.1056/NEJMra072149

[2] ACR: Radiation Dose to Adults From Common Imaging Examinations. https://www.acr.org/-/media/ACR/Files/Radiology-Safety/Radiation-Safety/Dose-Reference-Card.pdf

[3] null: Reduced lung-cancer mortality with low-dose computed tomographic screening. New England Journal of Medicine **365**(5), 395–409 (2011) https://doi.org/10.1056/NEJMoa1102873

[4] Chen, H., Zhang, Y., Zhang, W., Liao, P., Li, K., Zhou, J., Wang, G.: Low-dose CT via convolutional neural network. Biomedical Optics Express **8**(2), 679 (2017) https://doi.org/10.1364/BOE.8.000679

[5] Xia, W., Lyu, Q., Wang, G.: Low-dose CT using denoising diffusion probabilistic model for 20$\times$ speedup. CoRR **abs/2209.15136** (2022) https://doi.org/10.48550/ARXIV.2209.15136 2209.15136

[6] Yang, Q., Yan, P., Zhang, Y., Yu, H., Shi, Y., Mou, X., Kalra, M.K., Zhang, Y., Sun, L., Wang, G.: Low-dose CT image denoising using a generative adversarial network with wasserstein distance and perceptual loss. IEEE Transactions on Medical Imaging **37**(6), 1348–1357 (2018) https://doi.org/10.1109/TMI.2018.2827462

[7] Huang, B., Lu, S., Zhang, L., Lin, B., Wu, W., Liu, Q.: One-sample diffusion modeling in projection domain for low-dose CT imaging. IEEE Transactions on Radiation and Plasma Medical Sciences **8**(8), 902–915 (2024)

[8] Gao, Q., Shan, H.: CoCoDiff: a contextual conditional diffusion model for low-dose CT image denoising. In: Müller, B., Wang, G. (eds.) Developments in X-Ray Tomography XIV, p. 16. SPIE, San Diego, United States (2022). https://doi.org/10.1117/12.2634939

[9] Gao, Q., Li, Z., Zhang, J., Zhang, Y., Shan, H.: CoreDiff: Contextual Error-Modulated Generalized Diffusion Model for Low-Dose CT Denoising and Generalization. IEEE Transactions on Medical Imaging, 1–1 (2023) https://doi.org/10.1109/TMI.2023.3320812

[10] Ma, Y., Wei, B., Feng, P., He, P., Guo, X., Wang, G.: Low-dose CT image denoising using a generative adversarial network with a hybrid loss function for noise learning. IEEE Access **8**, 67519–67529 (2020) https://doi.org/10.1109/ACCESS.2020.2986388

[11] Kim, B., Han, M., Shim, H., Baek, J.: A performance comparison of convolutional neural network-based image denoising methods: The effect of loss functions on low-dose CT images. Medical Physics **46**(9), 3906–3923 (2019) https://doi.org/10.1002/mp.13713

[12] Li, M., Hsu, W., Xie, X., Cong, J., Gao, W.: SACNN: Self-attention convolutional neural network for low-dose CT denoising with self-supervised perceptual loss network. IEEE Transactions on Medical Imaging **39**(7), 2289–2301 (2020) https://doi.org/10.1109/TMI.2020.2968472

[13] Wang, R., Chen, Z., Song, Z., Fang, W., Zhang, J., Tu, D., Tang, Y., Xu, M., Ye, X., Lu, L., Jin, D.: Anatomy-Aware Low-Dose CT Denoising via Pretrained Vision Models and Semantic-Guided Contrastive Learning . In: Proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025,

vol. LNCS 15961. Springer, ??? (2025)

[14] Gong, G., Yu, M.: A Denoising Framework for Real-World Ultra-Low-Dose Lung CT Images Based on an Image Purification Strategy (2025)

[15] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., III, W.M.W., Frangi, A.F. (eds.) Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III. Lecture Notes in Computer Science, vol. 9351, pp. 234–241. Springer, ??? (2015). https://doi.org/10.1007/978-3-319-24574-4_28

[16] Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021) https://doi.org/10.1038/s41592-020-01008-z

[17] Zhang, X., Su, T., Zhang, Y., Cui, H., Tan, Y., Zhu, J., Xia, D., Zheng, H., Liang, D., Ge, Y.: Transferring u-net between low-dose CT denoising tasks: a validation study with varied spatial resolutions. Quantitative Imaging in Medicine and Surgery **14**(1), 640–652 (2024) https://doi.org/10.21037/qims-23-768

[18] Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: Dpm-solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022 (2022)

[19] Bushong, S.C.: Radiologic Science for Technologists E-book: Radiologic Science for Technologists E-book. Elsevier Health Sciences, ??? (2020)

[20] Lyu, P., Liu, N., Harrawood, B., Solomon, J., Wang, H., Chen, Y., Rigiroli, F., Ding, Y., Schwartz, F.R., Jiang, H., *et al.*: Is it possible to use low-dose deep learning reconstruction for the detection of liver metastases on ct routinely? European Radiology **33**(3), 1629–1640 (2023)

[21] Zeng, D., Huang, J., Bian, Z., Niu, S., Zhang, H., Feng, Q., Liang, Z., Ma, J.: A simple low-dose x-ray ct simulation from high-dose scan. IEEE transactions on nuclear science **62**(5), 2226–2233 (2015)

[22] Wasserthal, J., Breit, H.-C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., Bach, M., Segeroth, M.: Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. Radiology: Artificial Intelligence **5**(5), 230024 (2023)