DiffNR: Diffusion-Enhanced Neural Representation Optimization for Sparse-View 3D Tomographic Reconstruction

Shiyan Su[1*], **Ruyi Zha**[2*], **Danli Shi**[3], **Hongdong Li**[2], **Xuelian Cheng**[1†]
[1]Department of Data Science & Artificial Intelligence (DSAI), Monash University
[2]The Australian National University
[3]Hong Kong Polytechnic University
Xuelian.Cheng@monash.edu

Abstract

Neural representations (NRs), such as neural fields and 3D Gaussians, effectively model volumetric data in computed tomography (CT) but suffer from severe artifacts under sparse-view settings. To address this, we propose DiffNR, a novel framework that enhances NR optimization with diffusion priors. At its core is SliceFixer, a single-step diffusion model designed to correct artifacts in degraded slices. We integrate specialized conditioning layers into the network and develop tailored data curation strategies to support model finetuning. During reconstruction, SliceFixer periodically generates pseudo-reference volumes, providing auxiliary 3D perceptual supervision to fix underconstrained regions. Compared to prior methods that embed CT solvers into time-consuming iterative denoising, our repair-and-augment strategy avoids frequent diffusion model queries, leading to better runtime performance. Extensive experiments show that DiffNR improves PSNR by 3.99 dB on average, generalizes well across domains, and maintains efficient optimization.

Introduction

X-ray computed tomography (CT) is an essential imaging technique for noninvasive inspection of internal structures. A CT scanner captures multi-view projections that record the X-ray attenuation through the material. Given these projections, 3D tomographic reconstruction aims to recover a radiodensity volume. Conventional CT systems acquire hundreds of projections to produce a clean volume, but this results in substantial radiation exposure to subjects. Sparse-view CT (SVCT) reconstruction, which aims to maintain high-quality recovery with only a few dozen projections, thus becomes a crucial direction for safer imaging.

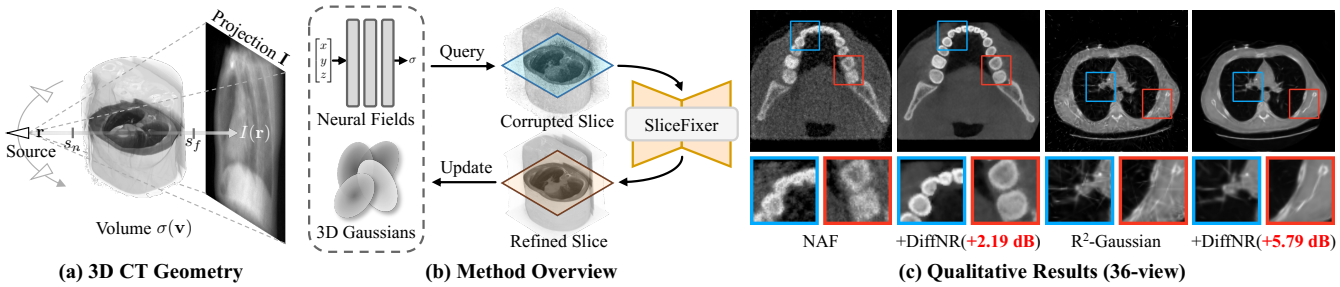
Recent years have seen rapid progress in learning-based SVCT. While feedforward approaches exist (Jin et al. 2017; Lin et al. 2024), optimization frameworks are generally preferred to enforce consistency between predicted volumes and measured projections. They can be broadly categorized into neural representation (NR) and neural prior (NP) approaches. NR methods model the volume as learnable neural fields (Zha, Zhang, and Li 2022) or 3D Gaussians (Zha et al. 2024), and optimize them in a self-supervised manner. They outperform traditional algorithms but yield artifacts in underconstrained regions. In contrast, NP methods pretrain networks to learn data-driven priors and then align network outputs with measurements using optimization solvers. Recent state-of-the-art NP approaches adopts unconditional 2D diffusion models (Ho, Jain, and Abbeel 2020) as network backbone, and embed local solvers into iterative denoising steps. While adequately steering unconditional generation towards the true data manifold, they suffer from inter-slice jitters, hallucinations, and long processing time.

In this work, we aim to marry neural representations with diffusion models. Unlike prior methods that embed local solvers into unconditional denoising processes, we adopt a fundamentally different strategy: enhancing a global NR with conditioned diffusion models. This design offers clear advantages: (1) learning a unified 3D representation promotes volumetric consistency, and (2) we can finetune powerful 2D foundation models instead of training one from scratch. Nevertheless, this integration is non-trivial, with key challenges in developing an NR-aware diffusion model and efficiently incorporating it into NR optimization.

To tackle these challenges, we propose DiffNR, **Diff**usion enhanced **N**eural **R**epresentation, for sparse-view 3D CT reconstruction. At its core is SliceFixer, a diffusion model specifically adapted to correct artifacts in NR-reconstructed slices. Leveraging 2D foundation models and recent advances in inference acceleration, we finetune a single-step diffusion model (Sauer et al. 2024) on a curated dataset of clean and corrupted slice pairs under varying sparsity levels. To improve structural awareness, we incorporate biplanar X-ray projections as additional conditioning inputs. During the reconstruction phase, SliceFixer periodically generates pseudo-reference volumes, which guide NR optimization in underconstrained regions. We adopt a perceptual SSIM-based regularization instead of voxel-wise losses to mitigate hallucinations and promote structural integrity. This repair-and-augment strategy reduces the need for frequent diffusion model queries, thus ensuring computational efficiency. We evaluate DiffNR across in-distribution and out-of-distribution datasets. Extensive experiments show that it improves NR reconstruction quality by 3.99 dB, generalizes well across domains, and maintains reasonable runtime.

We summarize our contributions as follows. (1) We pro-

*These authors contributed equally.
†Corresponding author: Xuelian Cheng



(a) 3D CT Geometry (b) Method Overview (c) Qualitative Results (36-view)

Figure 1: We propose DiffNR for sparse-view 3D CT reconstruction. (a) Geometry of a cone-beam CT scanner. (b) Method overview. (c) Comparison between the baseline methods (Zha, Zhang, and Li 2022; Zha et al. 2024) and our proposed DiffNR.

pose DiffNR, a novel framework that combines neural representation with diffusion priors, fundamentally different from prior CT methods. (2) We design an effective pipeline to adapt diffusion models for artifact correction and efficiently integrate them into NR optimization, which may also inspire other inverse problems. (3) Experiments demonstrate that DiffNR outperforms existing methods in accuracy, generalization, and efficiency, highlighting its practical values.

Related Work

Computed Tomography CT is widely used in daily applications such as medical diagnosis and security screening. Conventional fan-beam CT reconstructs a 3D volume slice by slice from 1D projection arrays. More recently, cone-beam CT has become popular as it swiftly captures 2D projection images, creating demand for direct volumetric reconstruction. Traditional algorithms fall into direct and iterative methods. Direct approaches (Feldkamp, Davis, and Kress 1984) instantly compute analytical results but produce severe artifacts. Iterative methods (Andersen and Kak 1984; Sidky and Pan 2008) formulate reconstruction as an optimization problem and solve it using numerical solvers. They reduce artifacts but oversmooth fine details.

Learning-Based Tomographic Reconstruction Similar to traditional algorithms, learning-based CT reconstruction can be performed directly or iteratively. Many works use feedforward networks to predict results from projections (Lin et al. 2024; Zhang et al. 2025) or low-quality reconstructions (Jin et al. 2017; Ma et al. 2023). Such a direct regression, however, lacks physical constraints. Consequently, more attention has shifted to optimization frameworks, broadly grouped into neural representation (NR) and neural prior (NP) approaches. NR methods, inspired by advances in RGB view synthesis such as NeRF (Mildenhall et al. 2020) and 3D Gaussian splatting (3DGS) (Kerbl et al. 2023), optimize a learnable field via differentiable rendering. There are NeRF (Zha, Zhang, and Li 2022; Cai et al. 2024) and 3DGS (Zha et al. 2024; Li et al. 2025) variants for 3D CT, but they struggle in sparse-view settings. NP methods combine optimization solvers (traditional or NR-based) with pretrained networks. Some methods use deterministic networks (Kamilov et al. 2023; Tian et al. 2025; Vo et al. 2024) as regularizer, and the state of the art plugs traditional local solvers into unconditional diffusion models (Chung et al. 2023; Chung, Lee, and Ye 2023). Within this paradigm, there are some early diffusion-NR hybrids (Du et al. 2024; Chu et al. 2025) which adapt NR as local solvers. Compared with prior methods, our DiffNR takes a new direction by enhancing a global NR with conditional diffusion models.

Diffusion-Enhanced Neural Representation Enhancing NR with diffusion priors has proven to be effective in RGB view synthesis. Some works use diffusion models as scorers that must be queried at each optimization step (Gu et al. 2023; Warburg et al. 2023; Zhou and Tulsiani 2023), which significantly compromises efficiency. Other approaches finetune diffusion models to repair corrupted images rendered from NR and augment training views with these pseudo-observations (Liu, Zhou, and Huang 2024; Liu et al. 2024). This strategy avoids frequent diffusion queries, thereby reducing computational overhead. Notably, Difix3D+ (Wu et al. 2025) further improves efficiency by employing single-step diffusion models (Sauer et al. 2024). Our method follows the repair-and-augment strategy but introduces key innovations designated for CT: (1) we correct artifacts on reconstructed slices rather than on rendered projections, and (2) we augment pseudo-volumes for direct 3D supervision instead of relying on intermediate image losses.

Background

X-ray Imaging This work adopts cone-beam geometry as a typical example of 3D CT, and the proposed method can be readily adapted to other geometries such as parallel-beam. As shown in Figure 1(a), an X-ray with initial intensity I_0 travels along the trajectory $\mathbf{r}(s) = \mathbf{o} + s\mathbf{d} \in \mathbb{R}^3$ where $s \in [s_n, s_f]$, passes through a density field $\sigma(\mathbf{v}) : \mathbb{R}^3 \to \mathbb{R}$ where $\mathbf{v}$ is any spatial location, and eventually reaches the detector plane. According to the Beer-Lambert law (Kak and Slaney 2001), the corresponding raw pixel value is given by $I'(\mathbf{r}) = I_0 \exp(-\int_{s_n}^{s_f} \sigma(\mathbf{r}(s))ds)$. In practice, raw data are transformed into logarithmic space for computational convenience, yielding the processed pixel value: $I(\mathbf{r}) = \log I_0 - \log I'(\mathbf{r}) = \int_{s_n}^{s_f} \sigma(\mathbf{r}(s))\, ds$. Unless otherwise stated, we use the logarithmic projections as inputs. The goal of tomographic reconstruction is to recover the underlying density field $\sigma(\mathbf{v})$, typically output as a discrete voxel grid $\mathbf{V} \in \mathbb{R}^{X \times Y \times Z}$, from multi-angle projections

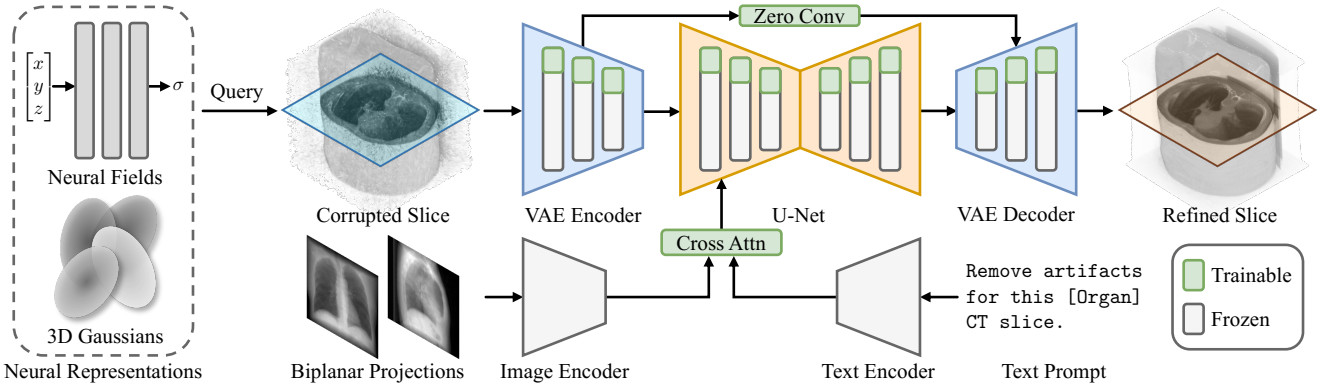


Figure 2: SliceFixer Architecture. It takes as input a CT slice queried from NRs, along with biplanar projections and a text prompt as conditions. It outputs a refined slice without artifacts. The model is built on SD-Turbo (Sauer et al. 2024), a single-step diffusion backbone. Trainable LoRA layers and zero convolutions are injected to adapt the model for our purpose.

$\{\mathbf{I}_i\}_{i=1}^N$. Note that real-world projections contain noise due to physical effects and hardware imperfections.

Neural Representations NR methods trains a 3D model via differentiable rendering. There are two primary types of NRs: neural fields and 3D Gaussians. Neural fields, as exemplified by NAF (Zha, Zhang, and Li 2022), represent the density field with a multilayer perceptron (MLP) f, which can be queried at any location $\mathbf{v}$ to produce the corresponding density $\sigma_f(\mathbf{v})$. The rendering function is a discrete Beer-Lambert law: $I_f(\mathbf{r}) = \sum_{i=1}^{P} \sigma_f(\mathbf{r}(s_i)) \cdot (\mathbf{r}(s_{i+1}) - \mathbf{r}(s_i))$ where P is the number of sampled points along each ray.

R^2-Gaussian (Zha et al. 2024) is a recent 3DGS-based approach, offering faster reconstruction than neural field methods. It represents the density field as a mixture of 3D Gaussians: $\sigma_g(\mathbf{v}) = \sum_{i=1}^{M} \mathcal{G}_i^3(\mathbf{v})$, where M is the number of kernels. Each Gaussian $\mathcal{G}_i^3$ has learnable parameters: base density ρ_i, center $\mathbf{p}_i \in \mathbb{R}^3$, and covariance $\mathbf{\Sigma}_i \in \mathbb{R}^{3\times3}$. Its form is given by: $\mathcal{G}_i^3(\mathbf{v}) = \rho_i \exp(-\frac{1}{2}(\mathbf{v}-\mathbf{p}_i)^\top \mathbf{\Sigma}_i^{-1}(\mathbf{v}-\mathbf{p}_i))$. To render a projection image, each 3D Gaussian is splatted onto the image plane as a 2D Gaussian $\mathcal{G}_i^2(\mathbf{u})$, where $\mathbf{u} \in \mathbb{R}^2$. The final projection is then computed by summing all 2D Gaussians: $I_g(\mathbf{u}) = \sum_{i=1}^{M} \mathcal{G}_i^2(\mathbf{u})$. We use NAF and R^2-Gaussian as two NR backbones.

Diffusion Models Diffusion models (Ho, Jain, and Abbeel 2020; Song et al. 2020) learn to approximate the data distribution p_{data} through iterative denoising. During training, a noisy version of a data sample $\mathbf{x} \sim p_{\text{data}}$ is generated as $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{x} + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$ is standard Gaussian noise, and $\bar{\alpha}_t$ controls noise level. The discrete diffusion timestep t is sampled from a uniform distribution $p_t \sim \mathcal{U}(0, t_{max})$. The denoising network θ predicts the added noise $\boldsymbol{\epsilon}_\theta$ and is optimized with the score matching objective: $\mathbb{E}_{\mathbf{x}\sim p_{\text{data}},\, t\sim p_t,\, \boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{1})}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t; \mathbf{c}, t)\|_2^2\right]$, where $\mathbf{c}$ denotes optional conditioning information, such as text or images. Recent advances (Sauer et al. 2024) accelerate diffusion inference by distilling the multi-step denoising process into a single-step generation.

Proposed Method

Given N projection images $\{\mathbf{I}_i\}_{i=1}^N$ acquired at uniform angular intervals around an object, our goal is to reconstruct its volumetric density field $\sigma(\mathbf{v})$, with emphasis on under-constrained regions that are prone to artifacts. To tackle this, we introduce DiffNR, a neural representation optimization framework with diffusion-based augmentation. This section is organized as follows. We begin by introducing SliceFixer, a single-step diffusion model that repairs degraded CT slices. Next, we detail the data curation strategies for model finetuning. Finally, we illustrate how to efficiently integrate SliceFixer into the optimization pipeline.

SliceFixer: Diffusion Model for Slice Repairing

Previous NR methods (Wu et al. 2025) repair artifacts at the projection level and incorporate intermediate image losses to optimize 3D models. While effective for surface-based RGB reconstruction, this strategy is suboptimal for volumetric reconstruction, where errors in penetrable X-ray projections accumulate. To address this, we propose SliceFixer, a diffusion model that predicts a refined slice $\hat{\mathbf{S}} \in \mathbb{R}^{X'\times Y'}$ from its counterpart $\tilde{\mathbf{S}}$ queried from NRs. We build SliceFixer upon SD-Turbo (Sauer et al. 2024), a single-step diffusion model that has demonstrated strong performance in image-to-image translation tasks (Parmar et al. 2024) and providing good inference efficiency. Following Chung et al. (2023), we use axial (z-direction) slices in practice, though the approach can be extended to arbitrary slicing directions. Architecture is shown in Figure 2. A VAE encodes corrupted slices into latents, and a U-Net predicts the target latents conditioned on the encoded inputs, conditions, and denoising timestep. The refined slice is then reconstructed using the VAE decoder.

Conditioning We aim to teach SliceFixer to remove artifacts while preserving anatomical structures in CT slices. To this end, our model is conditioned jointly on a text prompt c_t and two orthogonal X-ray projections $(\mathbf{I}_a, \mathbf{I}_b)$. The text prompt provides high-level semantic guidance, whereas the biplanar X-ray projections contains global structural cues.

Algorithm 1: Diffusion-Enhanced NR Optimization

Input: Sparse-view projections $\{\mathbf{I}_i\}_{i=1}^N$, scanner calibration parameters $\{\mathbf{K}_i\}_{i=1}^N$, neural fields f or 3D Gaussians g
Output: Density volume $\mathbf{V}$

for $j = 1$ to J **do**
  Render projection $\tilde{\mathbf{I}}_i$ with geometry parameters $\mathbf{K}_i$
  Compute L1 and SSIM losses between $\tilde{\mathbf{I}}_i$ and $\mathbf{I}_i$
  Query volume $\tilde{\mathbf{V}}_{tv}$ and compute total variation (TV)
  if $j \bmod \ell = 0$ **then**
    Query volume $\tilde{\mathbf{V}}_\ell$
    for each axial slice $\tilde{\mathbf{S}}$ in $\tilde{\mathbf{V}}_\ell$ **do**
      Upsample $\tilde{\mathbf{S}}$ to match `SliceFixer` input size
      Generate repaired slice $\hat{\mathbf{S}}$ with `SliceFixer`
      Downsample $\hat{\mathbf{S}}$ back to queried size
    end for
    Stack repaired slices into a volume $\hat{\mathbf{V}}_\ell$
  end if
  if $\hat{\mathbf{V}}_\ell$ exists and $j \bmod \tau = 0$ **then**
    Query $\tilde{\mathbf{V}}$ and compute its 3D SSIM loss with $\hat{\mathbf{V}}_\ell$
  end if
  Update f or g based on all losses
end for
Query final volume $\mathbf{V}$ from trained f or g

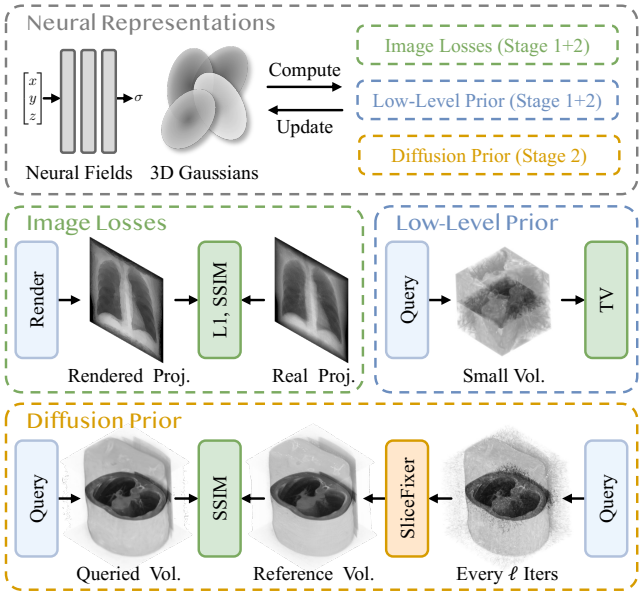


Figure 3: DiffNR Pipeline. During the training, we train neural representations using image losses and low-level regularization. In Stage 2, we generate a pseudo-reference volume with SliceFixer every ℓ iterations, and then apply SSIM regularization on queried and reference volumes.

We employ the pretrained RAD-DINO (Pérez-García et al. 2025) tailored for radiographs to encode image features. These image features are subsequently aggregated with text embedding via a cross-attention layer to form the conditioning input $\mathbf{c} = \texttt{Embed}(\mathbf{I}_a, \mathbf{I}_b, c_t)$ for the diffusion model.

Finetuning We finetune a pretrained 2D foundation model SD-Turbo (Sauer et al. 2024) to leverage its rich visual priors. Following Pix2pix-Turbo (Parmar et al. 2024), we inject LoRA adapters (Hu et al. 2022) into the VAE and U-Net modules and incorporate skip connections between the encoder and decoder via zero-convolution layers (Zhang, Rao, and Agrawala 2023). Other parameters are kept frozen.

Losses We integrates several standard diffusion losses, including L2 loss, LPIPS loss (Zhang et al. 2018), CLIP alignment loss (Radford et al. 2021), and an adversarial loss implemented with a CLIP-based discriminator for the target domain (Parmar et al. 2024). Additionally, we introduce a structural similarity (SSIM) (Wang et al. 2004) loss that captures perceptual quality. Our final objective is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{L2}} + \mathcal{L}_{\text{LPIPS}} + \lambda_{\text{CLIP}}\mathcal{L}_{\text{CLIP}} + \lambda_{\text{GAN}}\mathcal{L}_{\text{GAN}} + \lambda_{\text{SSIM}}\mathcal{L}_{\text{SSIM}}.$$

Data Curation

Training SliceFixer requires a large-scale dataset of paired slices, where one slice contains artifacts typically introduced during NR optimization and the other serves as the clean ground truth. However, no existing dataset satisfies these requirements. To address this, we leverage public 3D CT volumes to synthesize projection data and train a diverse set of neural representations. We explore various strategies to expand the training set and improve data diversity.

View Distribution We use the tomography toolbox (Biguri et al. 2016) to synthesize K dense projections for each real CT volume over a full 360° angular range. To simulate sparse-view scenarios, we randomly sample subsets of these projections to train NR models. We explore both uniformly and non-uniformly distributed view configurations. This variation introduces diverse artifact patterns in the reconstructed volumes, thereby enhancing the model's robustness to varying sparse-view conditions.

Model Underfitting We intentionally underfit the NR optimization by limiting training to a reduced number of iterations (e.g., 25–50% of the standard training steps). These underfitted reconstructions exhibit more pronounced artifacts due to incomplete convergence, thereby enriching the training set with challenging examples.

Mixed Neural Representation We mix reconstruction results from both neural fields and 3D Gaussians in a 1:1 ratio to encourage the diffusion model to learn generalized priors, rather than overfitting to specific patterns.

DiffNR: Diffusion-Enhanced Neural Representation Optimization

While SliceFixer effectively suppresses artifacts, it may introduce hallucinated details, which is highly undesirable in medical diagnostics. Moreover, this 2D model fails to maintain volumetric consistency, resulting in noticeable inter-slice jitters. To address these issues, instead of treating Slice-

Methods	ToothFairy (Cipriano et al. 2022)			LUNA16 (Setio et al. 2017)			TIME
	36-view PSNR / SSIM	24-view PSNR / SSIM	12-view PSNR / SSIM	36-view PSNR / SSIM	24-view PSNR / SSIM	12-view PSNR / SSIM	
Traditional Methods							
SART	27.41 / 0.581	27.13 / 0.596	25.66 / 0.604	22.34 / 0.438	21.77 / 0.437	19.96 / 0.412	1m25s
ASD-POCS	29.65 / 0.775	28.34 / 0.765	25.91 / 0.721	23.93 / 0.661	22.63 / 0.616	20.04 / 0.512	48s
Diffusion-Based Iterative Methods							
DiffusionMBIR	<u>33.29</u> / 0.856	30.54 / 0.818	26.28 / 0.733	**29.35** / 0.781	<u>27.15</u> / 0.735	23.01 / 0.581	11h15m
DDS	32.56 / 0.817	<u>31.13</u> / 0.788	<u>28.66</u> / 0.767	26.21 / 0.554	25.21 / 0.512	<u>23.29</u> / 0.486	16m17s
Neural Representation Methods							
SAX-NeRF	28.48 / 0.835	27.91 / 0.832	26.11 / 0.812	23.72 / 0.704	23.20 / 0.690	21.50 / 0.639	4h9m
NAF	28.62 / 0.833	28.20 / 0.833	26.22 / 0.812	23.85 / 0.712	23.18 / 0.692	21.37 / 0.618	7m15s
+DiffNR (Ours)	31.27 / **0.951**	30.79 / **0.946**	28.10 / **0.906**	26.27 / **0.867**	25.15 / **0.839**	22.98 / **0.765**	8m41s
R^2-Gaussian	28.56 / 0.695	26.36 / 0.634	22.63 / 0.537	24.11 / 0.577	22.06 / 0.497	18.32 / 0.364	5m52s
+DiffNR (Ours)	**33.52** / <u>0.900</u>	**32.92** / <u>0.895</u>	**29.71** / <u>0.852</u>	<u>28.82</u> / <u>0.822</u>	**27.43** / <u>0.793</u>	**24.37** / <u>0.712</u>	11min35s

Table 1: Quantitative results on ToothFairy and LUNA16 datasets. The best values are in bold, second-best are underlined.

Fixer as a post-processing module, we integrate it into the NR optimization process. DiffNR pipeline is illustrated in Figure 3, and the algorithm is shown in Algorithm 1.

Enhanced Volumes as Augmented Supervision We begin by optimizing a NR using standard image losses (L1 and SSIM) and low-level 3D regularization (total variation (Rudin, Osher, and Fatemi 1992)) to capture global structures. Every ℓ iterations, we query a volume $\tilde{\mathbf{V}}_\ell$ from the current model. We then upsample its slices using bilinear interpolation, apply SliceFixer for artifact correction, and downsample the results to the original resolution, producing a pseudo-reference volume $\hat{\mathbf{V}}_\ell$. We show in ablation that this up-downsampling strategy improves reconstruction quality. For the remaining training steps, we augment with an additional 3D supervision between the queried volume $\tilde{\mathbf{V}}$ and this reference volume $\hat{\mathbf{V}}_\ell$ every τ steps. This repair-and-augment strategy reduces the frequency of SliceFixer queries, thus preserving the overall optimization efficiency.

Perceptual Loss for Structural Integrity SliceFixer may introduce hallucinated details not perfectly aligned with measured projections. Consequently, directly minimizing voxel-wise L1 loss, as commonly adopted in image supervision, can lead to suboptimal performance. To address this, we adopt a perceptual loss based on 3D SSIM, computed as the average of 2D SSIM scores across axial, sagittal, and coronal planes. This promotes structural coherence and smoothness in underconstrained regions, rather than overfitting to fine-grained, potentially hallucinated details. We use a loss weight λ_{diff} to balance the contribution of 3D SSIM.

Experiments

Experimental Setup

Datasets We use two datasets: ToothFairy (Cipriano et al. 2022) and LUNA16 (Setio et al. 2017). ToothFairy consists of 443 dental scans, split into 393/25/25 for training/validation/testing, respectively. LUNA16 includes 888 chest scans, divided into 838/25/25. We train a separate SliceFixer on each dataset and apply the corresponding model for test-case reconstruction. We follow Lin et al. (2024); Zha, Zhang, and Li (2022) to preprocess raw CT volumes to a resolution of 256^3 and X-ray projections to 256^2. Sparse-view reconstruction is defined as using fewer than a hundred views, and we evaluate the challenging 36-, 24-, and 12-view settings.

Implementation Details SliceFixer is finetuned from SD-Turbo on 512^2 images, which are upsampled from 256^2 slices. We integrate LoRA layers with ranks of 8 for the U-Net and 4 for the VAE, and train the model with a learning rate of $1\text{e}{-5}$ for 40k steps on ToothFairy and 70k steps on LUNA16, using a batch size of 4. Loss weights are set to $\lambda_{\text{CLIP}} = 4$, $\lambda_{\text{GAN}} = 0.4$, and $\lambda_{\text{SSIM}} = 0.5$. Finetuning is performed on 4 H100 GPUs. DiffNR is implemented in PyTorch and optimized using the Adam optimizer (Kingma 2014). We use NAF and R^2-Gaussian as backbones, training them for 11k and 13.5k epochs, respectively, while keeping other hyperparameters unchanged. We empirically set $\ell = 10\text{k}$, and use $\tau = 20$ for NAF and $\tau = 10$ for R^2-Gaussian. Pseudo-reference volumes have a resolution of 256^3. All test-case reconstructions are performed on an RTX 3090 GPU. The code and model will be publicly available.

Compared Methods and Evaluation We compare with widely-used optimization-based methods, including (1) traditional iterative methods SART (Andersen and Kak 1984) and ASD-POCS (Sidky and Pan 2008), (2) self-supervised NR methods SAX-NeRF (Cai et al. 2024), NAF (Zha, Zhang, and Li 2022), and R^2-Gaussian (Zha et al. 2024), and (3) diffusion-based iterative methods: DDS (Chung, Lee, and Ye 2023) and DiffusionMBIR (Chung et al. 2023). We quantitatively evaluate all methods using standard metrics PSNR and SSIM.

Results

In-Distribution Performance Table 1 presents quantitative results on ToothFairy and LUNA16. Traditional methods and self-supervised NR approaches produce significant artifacts. While diffusion-based methods achieve higher

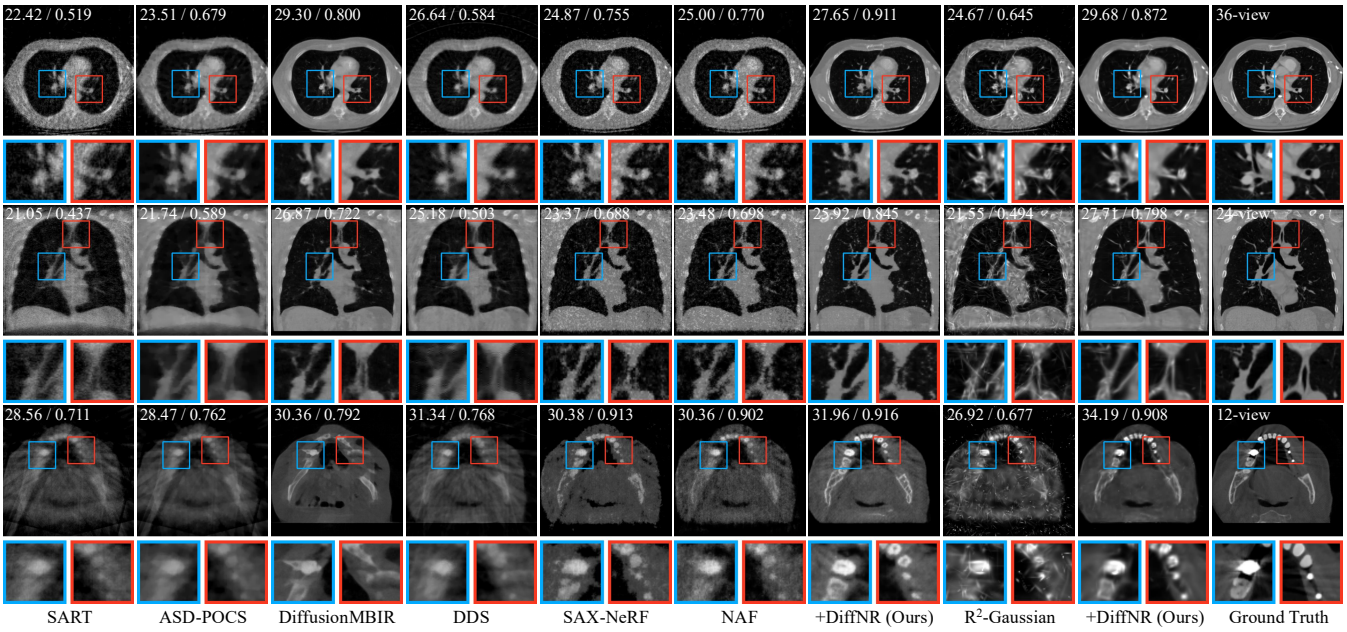


Figure 4: Qualitative results of reconstructed volumes on two datasets, shown from different slicing directions and sparsity levels. We annotate PSNR/SSIM on the top-left of each image. DiffNR recovers finer details and effectively suppresses artifacts.

Methods	OOD Dataset (Zha et al. 2024)		
	36-view PSNR / SSIM	24-view PSNR / SSIM	12-view PSNR / SSIM
SART	30.50 / 0.740	29.43 / 0.721	27.53 / 0.695
ASD-POCS	32.28 / 0.852	30.16 / 0.811	27.36 / 0.750
DiffusionMBIR	33.26 / 0.839	30.97 / 0.796	26.82 / 0.668
DDS	29.45 / 0.638	26.97 / 0.536	25.17 / 0.520
R^2-Gaussian	<u>35.64</u> / <u>0.904</u>	<u>33.46</u> / <u>0.868</u>	<u>29.71</u> / <u>0.792</u>
+DiffNR (Ours)	**35.99 / 0.918**	**34.15 / 0.896**	**31.04 / 0.848**

Table 2: Quantitative results on the OOD dataset. The best values are in bold, second-best are underlined

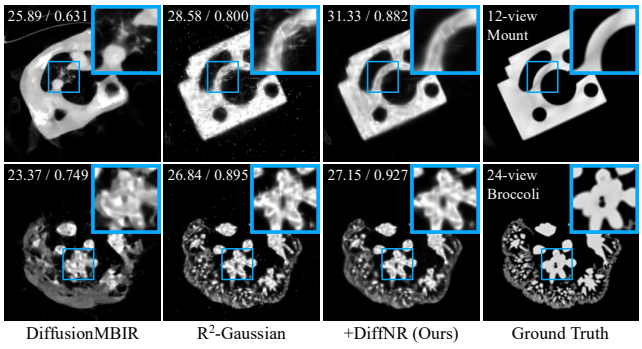


Figure 5: Qualitative results on OOD dataset.

scores, they come at the cost of hallucinated details and significant computation time. Previous SOTA DiffusionMBIR takes 11 hours to process a single case. In contrast, our DiffNR consistently enhances NR baselines, yielding an average improvement of +2.19 dB in PSNR for NAF and +5.79 dB for R^2-Gaussian. Although DiffNR introduces additional optimization time, it remains substantially faster than prior diffusion-based methods. Qualitative comparisons are provided in Figure 4, where DiffNR recovers fine structures and substantially reduces artifacts present in NR baselines.

Out-of-Distribution Performance To evaluate generalization capability, we use SliceFixer pretrained on ToothFairy and apply R^2-Gaussian+DiffNR to dataset from Zha et al. (2024), which includes 18 diverse cases spanning human organs, biological specimens, and artificial objects. Notably, this dataset contains real-world captured projections. Quantitative and qualitative results are shown in Table 2 and Figure 5, respectively. DiffNR outperforms other methods by suppressing hallucinations and artifacts, which shows that SliceFixer learns generalizable artifact patterns.

Downstream Application We further validate our method on downstream medical tasks such as segmentation. Specifically, we use the LungMask toolkit (Hofmanninger et al. 2020) to perform left/right lung segmentation on the reconstructed volumes. We use Dice (Dice 1945) and average surface distance (ASD) metrics to evaluate performance. As shown in Table 3 and Figure 6(a), the segmentation masks generated from Gaussian-based DiffNR are more consistent with those obtained from the ground-truth volumes, demonstrating the practical utility of our method.

Ablation Study

SliceFixer Design We validate design choices of SliceFixer in Table 4 and Figure 6(b). We find that finetuning

Methods	36-view Dice↑/ASD↓	24-view Dice↑/ASD↓	12-view Dice↑/ASD↓
SART	81.89 / 11.73	74.12 / 16.63	56.92 / 26.62
ASD-POCS	76.47 / 15.71	70.06 / 19.64	57.98 / 22.27
DiffusionMBIR	90.33 / 6.13	86.96 / 6.97	77.75 / 11.96
DDS	80.03 / 16.66	75.98 / 16.60	68.23 / 19.20
R^2-Gaussian	90.41 / 5.19	84.32 / 8.39	59.73 / 25.11
+DiffNR (Ours)	**93.74 / 3.85**	**90.71 / 5.60**	**84.93 / 9.59**

Table 3: Quantitative results for lung segmentation of reconstructed results on LUNA16 dataset.

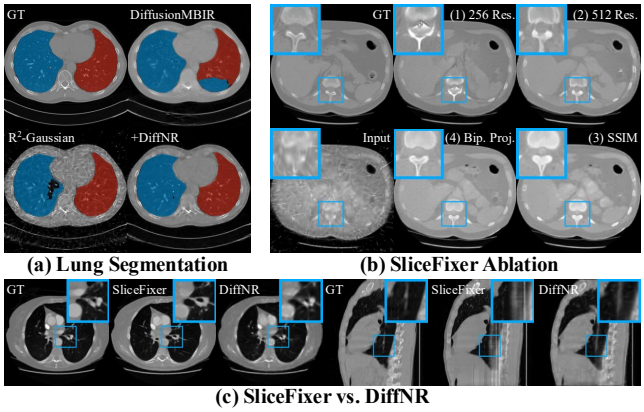
ID	Res.	SD-Turbo Pretrain	$\mathcal{L}_{ssim}$	Bip. Proj.	PSNR	SSIM
(1)	256	✓			27.65	0.789
(2)	512	✓			27.91	0.807
(3)	512	✓	✓		28.21	0.814
(4)	512	✓	✓	✓	**28.82**	**0.822**

Table 4: Ablation study of SliceFixer. We finetune different models and evaluate DiffNR under LUNA16 36-view cases.

SliceFixer on 512^2 images and applying up-downsampling to queried slices leads to better reconstruction quality compared to using the original 256^2 resolution. Additionally, incorporating an SSIM loss into finetuning resulting in a 0.3 dB gain in PSNR. Finally, adding biplanar projections as additional conditioning inputs provides rich structural cues and further boosts finetuning performance by 0.6 dB in PSNR.

DiffNR Design We use R^2-Gaussian as backbone to validate components of DiffNR as shown in Table 5. Augmenting NRs with novel-view images, commonly used in RGB surface reconstruction (Wu et al. 2025), is ineffective in volume reconstruction. This is because errors in penetrable projections can accumulate to the target volume across views. Instead, we choose to augment slice supervision, which proves to be more stable and effective. Moreover, applying SliceFixer as a standalone post-processing step leads to slice jitter and hallucinations (Figure 6(c)), highlighting the necessity of integrating it into the optimization pipeline. Lastly, we find that using voxel-wise L1 loss results in a performance drop, as the pseudo-reference volumes may contain details inconsistent with measured projections. A 3D perceptual loss is thus preferred. Overall, integrating our proposed components leads to the best performance.

Parameter Analysis We perform parameter analysis for Gaussian-based DiffNR to investigate the impact of 3D SSIM loss weight λ_{diff} and 3D supervision frequency τ. As shown in Table 6, $\lambda_{\text{diff}} = 0.5$ achieves the best performance by balancing the guidance from 3D supervision and avoiding overfitting to projections or degradation from diffusion hallucination. For the supervision interval, $\tau = 10$ yields optimal results. More frequent supervision (e.g., $\tau = 5$) may lead to over-reliance on the 3D loss and increased computational cost, whereas sparse supervision (e.g., $\tau = 20$) weakens structural regularization and degrades performance.



(a) Lung Segmentation **(b) SliceFixer Ablation**

(c) SliceFixer vs. DiffNR

Figure 6: Qualitative results of downstream tasks and ablation study. (a) Lung segmentation with the left lung in blue and the right lung in red. (b) Visualization of different design choices for SliceFixer. (c) Comparison of standalone SliceFixer post-processing and our proposed DiffNR.

Methods	PSNR	SSIM
R^2-Gaussian	24.11	0.577
+ Difix3D+ (augment projection)	23.23	0.579
+ SliceFixer (post-processing)	26.70	0.776
+ SliceFixer ($\mathcal{L}_1$)	26.42	0.678
+ SliceFixer ($\mathcal{L}_{ssim}$) (Ours)	**28.82**	**0.822**

Table 5: Ablation study of DiffNR design on LUNA16 dataset under 36-view setting.

λ_{diff}	0.3	**0.5**	0.7	1.0	1.5
PSNR	28.65	28.82	28.79	28.72	28.63
τ	5	**10**	15	20	30
PSNR	28.76	28.82	28.67	28.43	27.87
TIME	27m35s	12m56s	10m02s	8m32s	7m26s

Table 6: Ablation study of DiffNR hyperparameters on LUNA16 dataset (36-view) with our choices in bold.

Conclusion

We present DiffNR, a novel optimization framework for sparse-view 3D tomographic reconstruction. At its core is SliceFixer, a single-step diffusion model finetuned on curated datasets to correct artifacts in reconstructed CT slices. During reconstruction, the pretrained SliceFixer generates pseudo-reference volumes that provide augmented perceptual regularization. Such a repair-and-augment strategy avoids frequent diffusion model queries, therefore improving reconstruction quality without sacrificing efficiency. Experimental results demonstrate that DiffNR outperforms prior methods in reconstruction quality, generalization capability, and optimization efficiency, highlighting its practical potential. Further, this novel integration of diffusion models with neural representation optimization opens a promising direction for addressing broader classes of inverse problems.

Acknowledgments

This research is supported in part by the Jiangsu Department of Technology Natural Science Fund (Grants No: BK20250441), the Center of Excellence for Antimicrobial Therapeutics Discovery and Innovation (CEATDI, Grants No: 8002003), and the ARC Discovery Grant (Grant ID: DP220100800) of the Australia Research Council.

References

Andersen, A. H.; and Kak, A. C. 1984. Simultaneous algebraic reconstruction technique (SART): a superior implementation of the ART algorithm. *Ultrasonic imaging*, 6(1): 81–94.

Biguri, A.; Dosanjh, M.; Hancock, S.; and Soleimani, M. 2016. TIGRE: a MATLAB-GPU toolbox for CBCT image reconstruction. *Biomedical Physics & Engineering Express*, 2(5): 055010.

Cai, Y.; Wang, J.; Yuille, A.; Zhou, Z.; and Wang, A. 2024. Structure-aware sparse-view x-ray 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11174–11183.

Chu, J.; Du, C.; Lin, X.; Zhang, X.; Wang, L.; Zhang, Y.; and Wei, H. 2025. Highly accelerated MRI via implicit neural representation guided posterior sampling of diffusion models. *Medical Image Analysis*, 100: 103398.

Chung, H.; Lee, S.; and Ye, J. C. 2023. Decomposed diffusion sampler for accelerating large-scale inverse problems. *arXiv preprint arXiv:2303.05754*.

Chung, H.; Ryu, D.; McCann, M. T.; Klasky, M. L.; and Ye, J. C. 2023. Solving 3d inverse problems using pre-trained 2d diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22542–22551.

Cipriano, M.; Allegretti, S.; Bolelli, F.; Di Bartolomeo, M.; Pollastri, F.; Pellacani, A.; Minafra, P.; Anesi, A.; and Grana, C. 2022. Deep segmentation of the mandibular canal: a new 3D annotated dataset of CBCT volumes. *IEEE Access*, 10: 11500–11510.

Dice, L. R. 1945. Measures of the amount of ecologic association between species. *Ecology*, 26(3): 297–302.

Du, C.; Lin, X.; Wu, Q.; Tian, X.; Su, Y.; Luo, Z.; Zheng, R.; Chen, Y.; Wei, H.; Zhou, S. K.; et al. 2024. DPER: Diffusion prior driven neural representation for limited angle and sparse view CT reconstruction. *arXiv preprint arXiv:2404.17890*.

Feldkamp, L. A.; Davis, L. C.; and Kress, J. W. 1984. Practical cone-beam algorithm. *Josa a*, 1(6): 612–619.

Gu, J.; Trevithick, A.; Lin, K.-E.; Susskind, J. M.; Theobalt, C.; Liu, L.; and Ramamoorthi, R. 2023. Nerfdiff: Single-image view synthesis with nerf-guided distillation from 3d-aware diffusion. In *International Conference on Machine Learning*, 11808–11826. PMLR.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hofmanninger, J.; Prayer, F.; Pan, J.; Röhrich, S.; Prosch, H.; and Langs, G. 2020. Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. *European radiology experimental*, 4(1): 50.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Jin, K. H.; McCann, M. T.; Froustey, E.; and Unser, M. 2017. Deep convolutional neural network for inverse problems in imaging. *IEEE transactions on image processing*, 26(9): 4509–4522.

Kak, A. C.; and Slaney, M. 2001. *Principles of computerized tomographic imaging*. SIAM.

Kamilov, U. S.; Bouman, C. A.; Buzzard, G. T.; and Wohlberg, B. 2023. Plug-and-play methods for integrating physical and learned models in computational imaging: Theory, algorithms, and applications. *IEEE Signal Processing Magazine*, 40(1): 85–97.

Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4): 139–1.

Kingma, D. P. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Li, Y.; Fu, X.; Li, H.; Zhao, S.; Jin, R.; and Zhou, S. K. 2025. 3DGR-CT: Sparse-view CT reconstruction with a 3D Gaussian representation. *Medical Image Analysis*, 103585.

Lin, Y.; Yang, J.; Wang, H.; Ding, X.; Zhao, W.; and Li, X. 2024. Cˆ 2rv: Cross-regional and cross-view learning for sparse-view cbct reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11205–11214.

Liu, X.; Chen, J.; Kao, S.-H.; Tai, Y.-W.; and Tang, C.-K. 2024. Deceptive-NeRF/3DGS: Diffusion-Generated Pseudo-Observations for High-Quality Sparse-View Reconstruction. In *European Conference on Computer Vision*, 337–355. Springer.

Liu, X.; Zhou, C.; and Huang, S. 2024. 3dgs-enhancer: Enhancing unbounded 3d gaussian splatting with view-consistent 2d diffusion priors. *Advances in Neural Information Processing Systems*, 37: 133305–133327.

Ma, C.; Li, Z.; Zhang, J.; Zhang, Y.; and Shan, H. 2023. FreeSeed: Frequency-band-aware and self-guided network for sparse-view CT reconstruction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 250–259. Springer.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *ECCV*.

Parmar, G.; Park, T.; Narasimhan, S.; and Zhu, J.-Y. 2024. One-step image translation with text-to-image models. *arXiv preprint arXiv:2403.12036*.

Pérez-García, F.; Sharma, H.; Bond-Taylor, S.; Bouzid, K.; Salvatelli, V.; Ilse, M.; Bannur, S.; Castro, D. C.; Schwaighofer, A.; Lungren, M. P.; Wetscherek, M. T.;

Codella, N.; Hyland, S. L.; Alvarez-Valle, J.; and Oktay, O. 2025. Exploring scalable medical image encoders beyond text supervision. *Nature Machine Intelligence*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Rudin, L. I.; Osher, S.; and Fatemi, E. 1992. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4): 259–268.

Sauer, A.; Lorenz, D.; Blattmann, A.; and Rombach, R. 2024. Adversarial diffusion distillation. In *European Conference on Computer Vision*, 87–103. Springer.

Setio, A. A. A.; Traverso, A.; De Bel, T.; Berens, M. S.; Van Den Bogaard, C.; Cerello, P.; Chen, H.; Dou, Q.; Fantacci, M. E.; Geurts, B.; et al. 2017. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge. *Medical image analysis*, 42: 1–13.

Sidky, E. Y.; and Pan, X. 2008. Image reconstruction in circular cone-beam computed tomography by constrained, total-variation minimization. *Physics in Medicine & Biology*, 53(17): 4777.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Tian, X.; Chen, L.; Wu, Q.; Du, C.; Shi, J.; Wei, H.; and Zhang, Y. 2025. Unsupervised Self-Prior Embedding Neural Representation for Iterative Sparse-View CT Reconstruction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(7): 7383–7391.

Vo, R.; Escoda, J.; Vienne, C.; and Decencière, É. 2024. Neural Field Regularization by Denoising for 3D Sparse-View X-Ray Computed Tomography. In *2024 International Conference on 3D Vision (3DV)*, 1166–1176. IEEE.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Warburg, F.; Weber, E.; Tancik, M.; Holynski, A.; and Kanazawa, A. 2023. Nerfbusters: Removing ghostly artifacts from casually captured nerfs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 18120–18130.

Wu, J. Z.; Zhang, Y.; Turki, H.; Ren, X.; Gao, J.; Shou, M. Z.; Fidler, S.; Gojcic, Z.; and Ling, H. 2025. Difix3d+: Improving 3d reconstructions with single-step diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 26024–26035.

Zha, R.; Lin, T. J.; Cai, Y.; Cao, J.; Zhang, Y.; and Li, H. 2024. R^2-Gaussian: Rectifying Radiative Gaussian Splatting for Tomographic Reconstruction. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Zha, R.; Zhang, Y.; and Li, H. 2022. NAF: neural attenuation fields for sparse-view CBCT reconstruction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 442–452. Springer.

Zhang, G.; Zha, R.; He, H.; Liang, Y.; Yuille, A.; Li, H.; and Cai, Y. 2025. X-lrm: X-ray large reconstruction model for extremely sparse-view computed tomography recovery in one second. *arXiv preprint arXiv:2503.06382*.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhou, Z.; and Tulsiani, S. 2023. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12588–12597.