MInDI-3D: Iterative Deep Learning in 3D for Sparse-view Cone Beam Computed Tomography

Daniel Barco[1], Marc Stadelmann[1], Martin Oswald[1], Ivo Herzig[2], Lukas Lichtensteiger[2], Pascal Paysan[3], Igor Peterlik[3], Michal Walczak[3], Bjoern Menze[4], and Frank-Peter Schilling[1]

[1]Centre for Artificial Intelligence (CAI), Zurich University of Applied Sciences (ZHAW), Winterthur, Switzerland.
[2]Institute of Applied Mathematics and Physics (IAMP), Zurich University of Applied Sciences (ZHAW), Winterthur, Switzerland.
[3]Varian Medical Systems Imaging Lab, Baden, Switzerland.
[4]Biomedical Image Analysis and Machine Learning, University of Zurich, Zurich, Switzerland.

August 14, 2025

Abstract

We present **MInDI-3D** (**M**edical **In**version by **D**irect **I**teration in **3D**), the first 3D conditional diffusion-based model for real-world sparse-view Cone Beam Computed Tomography (CBCT) artefact removal, aiming to reduce imaging radiation exposure. A key contribution is extending the "InDI" concept from 2D to a full 3D volumetric approach for medical images, implementing an iterative denoising process that refines the CBCT volume directly from sparse-view input. A further contribution is the generation of a large pseudo-CBCT dataset (16,182) from chest CT volumes of the CT-RATE public dataset to robustly train MInDI-3D. We performed a comprehensive evaluation, including quantitative metrics, scalability analysis, generalisation tests, and a clinical assessment by 11 clinicians. Our results show MInDI-3D's effectiveness, achieving a 12.96 (6.10) dB PSNR gain over uncorrected scans with only 50 projections on the CT-RATE pseudo-CBCT (independent real-world) test set and enabling an 8x reduction in imaging radiation exposure. We demonstrate its scalability by showing that performance improves with more training data. Importantly, MInDI-3D matches the performance of a 3D U-Net on real-world scans from 16 cancer patients across distortion and task-based metrics. It also generalises to new CBCT scanner geometries. Clinicians rated our model as sufficient for patient positioning across all anatomical sites and found it preserved lung tumour boundaries well.

1 Introduction

Reconstructing high-quality medical images from sparsely sampled or partial measurements is essential for advancing clinical imaging modalities such as computed tomography (CT), positron emission tomography (PET), and magnetic resonance imaging (MRI). These advancements aim to reduce scan times and patient radiation exposure. Among these modalities, cone beam computed tomography (CBCT) ex-

emplifies both the promise and challenges of sparse sampling.

CBCT is widely used to acquire volumetric X-ray images on radiation therapy treatment devices, such as linear accelerators, in image-guided radiation therapy [1]. It is also employed in interventional radiology offering high spatial resolution and short scan durations [2]. While pre-treatment planning CT offers higher image resolution for the intervention planning, the image of the day is acquired using on-device CBCT. These CBCT scans can enable tumour and organs at risk contouring, dose calculation, ART (adaptive radiation therapy) workflows and precise patient positioning [3]. Its clinical use faces the following challenges: First, image quality is often degraded by artefacts from patient motion, metal implants, and undersampled projections [4]. In addition, repeated daily scans over extended treatment periods (up to 40 sessions) raise concerns about cumulative radiation exposure of the patient.

To address the challenges of cumulative radiation exposure, reducing the number of projections, i.e. sparse-view CBCT, has been proposed. Sparse-view CBCT reconstruction however, introduces streak artefacts – due to the NyquistShannon sampling theorem being violated – which degrade image quality and hinder clinical utility. Deep learning-based approaches have emerged as promising solutions to address these challenges, offering the potential to reconstruct high-quality images from limited projection data.

Deep learning, particularly models including convolutional neural networks (CNNs), excels at learning hierarchical features from image data, making them well-suited for tasks such as image classification [5], segmentation [6], and reconstruction [7]. In medical imaging, U-Net [8] has become a cornerstone architecture due to its encoder-decoder structure with skip connections, which enables precise localisation and segmentation of anatomical structures. U-Net's success has inspired numerous variants and extensions, including 3D U-Net for volumetric data, which is particularly relevant for CBCT reconstruction.

Generative deep learning, including Generative Adversarial Networks (GANs) [9], Variational Autoencoders (VAEs) [10], and diffusion models [11], have enabled high quality unconditional image synthesis [12,13]. These unconditional generative models have been further extended into conditional image-to-image frameworks, where an input image is transformed into a desired output (e.g., an artefact-free, motion deblurred or super resolved image). Pioneering works like Pix2Pix [14] and CycleGAN [15] established architectures for paired and unpaired image-to-image translation, respectively. These methods have been adapted to medical imaging for tasks such as image-to-image domain translation (e.g., MRI to CT), image generation, segmentation and denoising/reconstruction (e.g. artefact removal) [16].

Diffusion models have significantly advanced image synthesis and restoration, surpassing traditional GANs in conditional and unconditional generation tasks [12,17,18]. Their iterative denoising process enables high-fidelity reconstructions by modelling complex data distributions. However, a key limitation of standard diffusion models is their computational cost and speed, as they often require hundreds of iterative steps during inference [11, 12]. This makes them impractical for many real-world applications, due to prolonged inference times. To address this, InDI (Inversion by Direct Iteration) [19] was proposed as an efficient alternative for image enhancement tasks. InDI reduces the required steps to a fraction by replacing the stochastic reverse diffusion process with a deterministic direct iteration approach. This approach achieves results comparable to traditional diffusion models with significantly fewer computational resources. However, so far, InDI has only been applied to 2D images and non-medical datasets. While classical diffusion models have shown promise in 3D medical image enhancement, their inherent computational cost remains a significant bottleneck for clinical adoption. Generalising efficient 2D frameworks to complex 3D volumetric data and inverse problems such as sparse-view CBCT introduces considerable technical challenges [20–23]. Thus, our work introduces MInDI-3D, a novel extension of InDI to 3D, which represents a key contribution for enabling efficient high-fidelity, volumetric medical image reconstruction. This gap in the literature motivates our work, which extends InDI to 3D and evaluates its performance in the context of sparse-view CBCT arte-

fact removal.

Our study compares the performance of MInDI-3D with a 3D U-Net, both almost identical in backbone-architecture and parameter count (time-embedding is added to the InDI backbone), highlighting the differences in their training strategies. We evaluate the impact of varying training data set sizes and different number of sparse-projections (25 and 50 projections, out of 400 projections in total for the test dataset) on the performance of these approaches. Extensive validation is conducted on test datasets acquired by a different scanner. The perception-distortion trade-off describes the inherent balance in image restoration tasks between achieving high perceptual quality (how "realistic" an image appears to a human observer) and minimising distortion (pixel-level deviations from the original) [24]. InDI enables control over this trade-off without retraining: increasing the amount of sampling steps, InDI can trade distortion for better perception reducing the problem of regression to the mean by adding realistic features [19]. We explore this perception-distortion trade-off.

Our main contributions are summarised as follows:

- We introduce MInDI-3D, the first fully 3D iterative diffusion-based model for sparse-view CBCT artefact removal, extending the 2D InDI concept to full 3D medical volumes.
- We generate and provide a large pseudo-CBCT dataset with 16,182 chest CT volumes including projections, enabling robust training of MInDI-3D.
- We conduct a comprehensive evaluation including quantitative metrics, scalability analysis, generalisation tests, and a clinical evaluation by 11 clinicians.
- Our results demonstrate MInDI-3D's effectiveness in achieving significant PSNR gains, enabling radiation exposure reduction, and showing strong generalisation to real-world data and new scanner geometries.

The remainder of this article is structured as follows: In the next subsection, we provide an overview of related work. In section 2, we discuss the datasets, data simulation, and deep learning methods used, elaborating on our training process and the metrics we used to evaluate our models. The following section 3 presents our findings across different datasets and model architectures, while also presenting the results of a clinical evaluation. Finally, in section 4 we discuss the results and provide an outlook on future research.

1.1 Related Work

Extensive research has been conducted on characterising and mitigating artefacts that degrade image quality in CT and CBCT reconstruction [25, 26]. In recent years, deep learning models have successfully been shown to reduce artefacts in both 3D and 4D (time-resolved) CBCT [4], offering promising solutions for enhancing sparse-view CBCT image quality (e.g. [27] for mitigation of motion artefacts). While numerous studies have explored artefact removal in sparse-view CBCT using deep learning, the majority of these approaches have focused on non-generative methods, often employing 2D approaches at times with spatial awareness to reduce computational complexity [28–30]. This spatial compromise creates an opportunity for fully 3D approaches, that by design optimise for inter-slice consistency.

Generative deep learning models in 3D have gained attention in the field of medical imaging for tasks such as unconditional image generation, image-to-image translation (e.g., MRI-to-CT), and image enhancement. Unconditional image generation has been proposed as a privacypreserving tool to augment small medical image datasets [31]. Three main architecture types have been used in 3D unconditional medical image generation: GANs [32, 33], VAEs [34, 35] and diffusion models [31, 36]. These developments in unconditional generation have naturally extended to conditional tasks requiring paired data. Image-to-image translation using generative models in 3D has shown impressive results for medical images [37–40]. Several studies were conducted using GAN-based approaches, while more recently, researchers have used diffusion and latent diffusion models for medical image to image tasks [41].

For medical image enhancement generative approaches have seen growing interest, though these approaches remain constrained by computational and practical challenges. While GAN-based approaches dominated early work [39, 42, 43], recent efforts have shifted toward diffusion-based approaches. Li et al. [21] employ three planar 2D diffusion models combined in iterative reconstruction with a measurement loss. They use the AAPM Low Dose CT Grand Challenge [44] dataset with 9 volumes for training and 1 for testing with a dimension of $512 \times 512 \times 512$. Lee et al. [22] use the same train and test dataset setup but use two perpendicular 2D diffusion models as a 3D prior. Li et al. [45] utilised a 2D score-based diffusion model for unconditional CT generation as a prior, combining it with a measurement loss on the CT Lymph Nodes Dataset [46] (156 subjects, 512×512 images). Most work to date has focused on 2D or pseudo-3D strategies leveraging triplane embeddings [20], 2.5D fusion [21], or separate 2D models [22,23] to manage the computational burden of 3D data [41]. These limitations have motivated research into efficient diffusion-based implementations. InDI requires a fraction of the steps compared to other diffusion-based models for the conditional setting and is therefore especially promising for the clinical setting, where reconstruction speed is essential [19].

2 Materials & Methods

2.1 Datasets

CT-RATE is a public dataset [47] that includes 25,692 non-contrast chest CT volumes, expanded to 50,188 through various reconstructions, from 21,304 unique patients Table 1. From this dataset, we use a subset of 3,612 patients. Volumes were of size 512×512 voxels in the transverse plane and on average 309 slices along the z-axis and an average spacing of $0.72 \times 0.72 \times 1$ mm on the x, y and z-axis. We used the CT-Rate dataset to generate a pseudo-CBCT training dataset. We forward-projected the CT volumes using a CBCT geometry to obtain CBCT projections (see section 2.2), which can then be reconstructed by a CBCT reconstruction algorithm to mimic the CBCT acquisition. Our pseudo-CBCT dataset – including projection images, sparse-view reconstructions (with 25, 50, and 100 projections), and corresponding ground truth volumes – is publicly available on Zenodo[1].

We used a real-world CBCT dataset for testing Table 1. This dataset was obtained in a Varian sponsored HyperSight imaging study (acquired on Varian Halcyon linear accelerators). We refer to this dataset as HyperSight. It comprises images from 16 cancer patients including five with abdominal cancer, five with breast cancer, and six with lung cancer, for whom permission to use their data has been obtained.

2.2 CBCT reconstruction and simulation

Reconstructing 3D CBCT volumes from 2D projections can be achieved through analytical and iterative approaches. The Feldkamp-Davis-Kress (FDK) [48] algorithm, an analytical method, provides a fast and reliable approximation of the inverse Radon transform, establishing itself as widely used baseline for 3D CBCT reconstruction. While FDK excels in computational efficiency, iterative reconstruction techniques – such as the Simultaneous Algebraic Reconstruction Technique (SART) [49] – leverage statistical models and iterative optimisation to improve image quality, particularly in sparse-view or low-dose scenarios. However, their high computational demands often render analytical methods like FDK more practical for routine clinical applications. Our implementation employs FDK with the Ram-Lak filter [50] to correct radial sampling non-uniformity, a method commonly termed filtered back-projection (FBP).

While the real-world dataset was acquired using a full-fan, half-trajectory geometry, the pseudo-CBCT was processed with a half-fan, full-trajectory scanning geometry. The full-trajectory configuration involves a 360° rotation, while the half-trajectory rotates 210°. Half-fan mode allows for a larger field of view by offsetting the detector laterally by 175 mm and using the entire detector for half the field of view. To mitigate artefacts from data redundancy in the

[1] https://zenodo.org/records/XXXXXXXX

Dataset	# Volumes	# Patients	Anatomic Region	Data Type	Scanner
CT-RATE	16182	3612	chest	pseudo-CBCT	Siemens SOMATOM
HyperSight	16	16	abdomen, breast, lung	CBCT	Varian Halcyon

Table 1: Dataset characteristics showing the pseudo-CBCT training dataset (with both volumes reconstructed with 491 and 697 projections) derived from 8091 CT chest scans (CT-Rate) enabling robust training and 16 real CBCT scans (HyperSight) validating clinical utility across multiple anatomic sites.

overlapping regions of the half-fan geometry, half-fan weighting was applied. The effective area of the real-world detector is 86×43 cm (3072×384 pixels). All projections were generated with a source-to-imager distance (SID) of 1540 mm and a source-to-axis distance (SAD) of 1000 mm.

For the pseudo-CBCT generation, CT volumes were forward-projected to simulate both full-view and sparse-view acquisitions. Projection parameters – detector size (366×160 pixels), pixel resolution (1.176×2.688 mm in axial and longitudinal directions, respectively), and projection counts (491 and 697 for full-view) – were aligned with a real-world half fan scan protocol from a Varian Halcyon machine. Reconstructions were performed using the FBP method, while varying the number of projections (full, 25, or 50). In sparse-view cases, projections were selected to uniform angular spacing, minimising clustering artefacts and ensuring optimal sampling coverage. The reconstructed volumes have a height, width and depth of $256 \times 256 \times 64$ voxels and a spacing of $2 \times 2 \times 3$ mm. We chose this volume size and spacing to balance memory constraints in our 3D deep learning pipelines with anatomical coverage.

2.3 Deep Learning Methods

This section presents the core methodology for correcting sparse-view artefacts in CBCT images using deep learning. First we present the architecture of our backbone U-Net [8, 51] (see Figure 1) and then proceed to the training and inference of MInDI-3D. Unlike many 3D based methods that resort to latent space or explicit spatial compression techniques like wavelets to mitigate memory challenges in volumetric data, our MInDI-3D operates directly in the 3D voxel space to preserve anatomical detail and reduce complexity.

2.3.1 3D U-Net backbone

Encoder Blocks: The encoder comprises four hierarchical stages. Each stage contains two residual submodules followed by downsampling. The first stage contains an additional input layer (kernel: $3 \times 3 \times 3$, stride: 1). The residual submodules process the volume as follows: (1) batch normalisation [52] (BN), (2) SiLU activation [53], and (3) a 3D convolution (kernel: $3 \times 3 \times 3$, stride: 1). The input to the residual submodule is then added to the output. After the residual blocks, a strided convolution (kernel: $3 \times 3 \times 3$, stride: 2) downsamples the feature map by a factor of 2. A skip connection adds the stage's output as input to the decoder at the same hierarchical level. Channel dimensions double at each stage, progressing from 32 to 512.

Decoder Blocks: The decoder mirrors the encoder, restoring spatial resolution through four stages. Each stage begins by concatenating the skip connection and the output from the lower stage and processing it with a residual submodule described above. The output is then upsampled with a transposed 3D convolution (kernel: $4 \times 4 \times 4$, stride: 2). Finally, on the last stage, an additional 3D convolutional layer (kernel: $3 \times 3 \times 3$, stride: 1) is employed. Channel dimensions halve at each stage, reversing the encoders progression (512 to 32).

Attention Mechanism: Convolutional attention applies the Scaled Dot-Product Attention [54] to a convolutional layer following [55]. The convolutional attention mechanism is integrated into the deepest two encoder and decoder layers and is described sub-

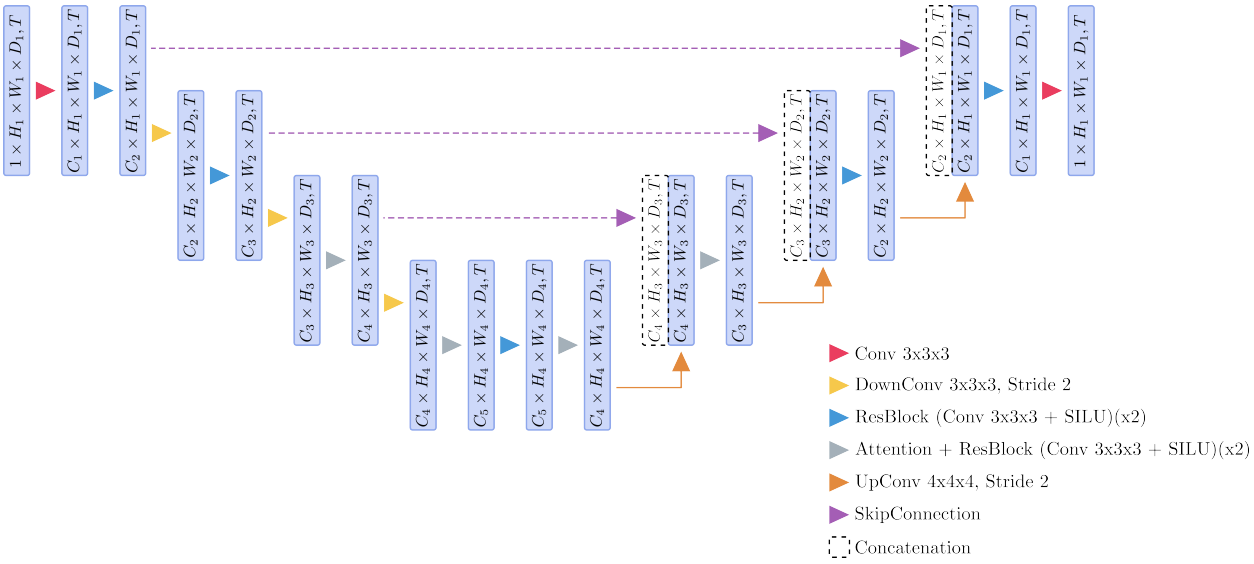

Figure 1: U-Net architecture with 4 hierarchical levels, showing layer-specific dimensionality ($C \times H \times W \times D$), where C is the number of channels, H is height, W is width, and D is depth (all in voxels), and time-embedding (T). SiLU (Sigmoid Linear Unit) activations introduce non-linearity.

sequently. Input features first undergo group normalisation, followed by a $1 \times 1 \times 1$ convolutions that project the normalised features into query, key, and value tensors. Attention weights are computed via scaled dot-product interactions across all spatial positions in the feature maps, enabling each voxel to dynamically aggregate information from the entire input domain. This global interaction is made tractable by applying the mechanism exclusively at deeper network stages, where hierarchical downsampling has reduced spatial dimensions.

2.3.2 Inversion by Direct Iteration (InDI)

InDI is a supervised image restoration method that avoids the "regression to the mean" effect, which can lead to over-correction of outputs toward the average of the training data. By gradually enhancing image quality in incremental steps, InDI produces more realistic and detailed images [19]. Unlike generative denoising diffusion models, InDI defines the restoration process directly from low-quality to high-quality image, and uses a convex combination of the input/target image as intermediate steps.

InDI forward degradation process: The InDI forward degradation process is defined as follows:

$$x_t = (1-t)x + ty, \quad \text{with} \quad t \in [0,1]. \tag{1}$$

x_t is an intermediate-degraded image between the low-quality input y (at $t = 1$) and the high-quality target x (at $t = 0$). The process starts from a clean image at $t = 0$ and degrades it to a noisy image at $t = 1$. The iterative restoration process then gradually improves the image quality by moving from $t = 1$ to $t = 0$ in small steps.

Iterative Restoration Process: The restoration phase inverts the forward process by iteratively predicting "cleaner" images while progressing backward from $t = 1$ to $t = 0$.

$$\hat{x}_{t-\frac{1}{N}} = \frac{1}{N \cdot t}\mathcal{F}_\theta\left(\hat{x}_t, t\right) + \left(1 - \frac{1}{N \cdot t}\right)\hat{x}_t \tag{2}$$

Equation (2) is a recursive update rule from the InDI framework, designed to refine a prediction iter-

atively. The left-hand side, $\hat{x}_{t-\frac{1}{N}}$, represents the next predicted time step, with N representing the number of steps. The right-hand side combines two terms: $\frac{1}{N \cdot t}\mathcal{F}_\theta\left(\hat{x}_t, t\right)$, which introduces a time-aware backbone model $\mathcal{F}_\theta$. This backbone model predicts the clean image from any time step/ degradation level. $\left(1-\frac{1}{N \cdot t}\right)\hat{x}_t$ accumulates the current estimate. As time progresses, the influence of the forward model diminishes, giving more weight to the accumulated estimate, ensuring stability.

In contrast to the baseline U-Net, we incorporate a time-embedding into the U-Net backbone of the InDI model. This time-embedding allows the model to understand the progression from the low-quality image to the high-quality image, effectively encoding the temporal distance between them and enabling an iterative restoration process. We use a sinusoidal time embedding proposed by [11] with 1024 channels.

2.4 Training

Training is conducted on an NVIDIA H200 GPU with 140 GB of VRAM, using the Adam optimiser [56] (learning rate 0.0001) and mean absolute error (MAE) (cf. 2.5) as the loss function. To improve convergence, we employ a learning rate scheduler (epoch step size 10, decay factor $\delta = 0.95$), a batch size of 4, and gradient accumulation every two steps. Models are trained for 500 epochs, taking approximately 57 hours when using a dataset with 320 subjects for training and 64 subjects for testing. The model with 3200 subjects took 216 hours to train for 180 epochs and was stopped thereafter due to time constraints (412 subjects were used for validation). We optimised the learning rate, learning rate scheduler, batch size, U-Net depth, size and attention layers for optimal performance. Input images were normalised by linearly mapping HU values from -1500 to 1000 onto a range from -1 to 1, without clipping.

2.5 Metrics & task-based Evaluation

In our experiments, we evaluate numerical distortion performance using several quantitative metrics that measure the point-wise voxel distance between pairs of images (x, x'):

- Mean absolute error MAE $= \frac{1}{N}\sum_{i=1}^{N}|x_i - x_i'|$, where N is the total number of images, x_i denotes the ground truth voxel value, and x_i' represents the predicted value;
- Structural Similarity Index Measure (SSIM) [57];
- Peak Signal-to-Noise Ratio PSNR $= 10\log_{10}\left(\frac{\text{MAX}^2}{\frac{1}{N}\sum_{i=1}^{N}(x_i-y_i)^2}\right)$, where MAX is the maximum possible pixel or voxel value;
- Dice Similarity Coefficient (DICE), which measures the spatial overlap between two segmented volumes, defined as $\text{Dice}(A,B) = \frac{2|A\cap B|}{|A|+|B|}$, where A and B denote the sets of voxels in the two segmentations. A Dice score of 1 indicates perfect overlap, while 0 indicates no overlap.

In Table 2, Table 3, Table 4, and Table 5, the standard deviation is shown after the mean of the metric. All distortion metrics are calculated in Hounsfield units (HU) from pairs of uncorrected or corrected volumes and their corresponding ground truth counterparts. SSIM quantifies structural similarity within spatially correlated 2D/3D regions. Flattening masked data into 1D arrays destroys these spatial relationships, rendering SSIM invalid for masked vectors. MAE and PSNR on the other hand measure pixel/voxel-wise errors, making them suitable for computation on flattened data. They are calculated exclusively for the body, with the air around the body masked out. The masks were generated by first applying Otsu's thresholding [58] method, which automatically determines an optimal threshold value to separate the foreground (typically the region of interest) from the background based on the image histogram. This binary segmentation was then refined using morphological operations. Dilation was used to close small gaps and connect nearby regions, while erosion helped remove small noise and further define the boundaries of the segmented structures. These metrics are referred to as masked. We calculate the PSNR value using 2000 HU as MAX value, corresponding to a range from -1000 to 1000 HU.

As perception metrics, we use the Frchet Distance (FD) [59] to measure the distance between the distribution of the ground truth compared to the predicted

image features. The image features are extracted using the pre-trained model DINOv2 [60], which as feature extractor has shown to most closely align with human perception [59]. We process the volumes along the axial plane and use the centre slice (2D image) as input for the DINOv2 encoder.

As task-based evaluation, we use TotalSegmentator [61] to segment the heart, left lung, right lung, ribs, and vertebrae. We chose TotalSegmentator as a segmentation tool due to its robust segmentation capabilities.

3 Results

3.1 Quantitative Evaluation

We evaluated the performance of the baseline 3D U-Net and the MInDI-3D models across multiple experimental configurations. Table 2 compares our MInDI-3D model against a 3D U-Net baseline. MInDI-3D achieves better results on both the hold-out test set from CT-RATE as well as the real-world HyperSight data (MAE: 30.55 vs. 32.16; PSNR 33.61 dB vs. 32.98 dB; SSIM: 0.91 vs. 0.90). The MInDI-3D model improved the volume reconstructed with 50 projections by $\Delta_{\text{PSNR}} = +12.96$ dB on the validation set and $\Delta_{\text{PSNR}} = +6.10$ dB on the test set, see Figure 2 .

To assess robustness to various levels of sparse-view inputs, we trained MInDI-3D with varying projection levels (25, 50, 100) and evaluated on the HyperSight dataset (Table 3). The image reconstructed with the smallest number of projections (sparse 25) achieved the largest relative improvement($\Delta_{\text{PSNR}} = +7.78$ dB), compared to the ground truth, while models trained with 100 projections showed the best absolute result (PSNR = 35.32 dB).

We perform a task-based evaluation of segmentation stability using TotalSegmentator to automatically categorise anatomical structures with MInDI-3D, comparing its performance to a 3D U-Net (Table 4). We evaluate how varying the number of iterative refinement steps (130) in MInDI-3D impacts segmentation accuracy, demonstrating that critical structures like lungs (DICE=0.960.99), vertebrae (DICE=0.95) and heart (DICE=0.910.92) are preserved and consistency is retained regardless of step count.

We present an ablation study on the performance of three MInDI-3D models, trained with no, 1, or 2 attention blocks (Table 5). Adding two attention blocks yielded $\Delta_{\text{MAE}} = -5.03$, $\Delta_{\text{PSNR}} = +2.02$ dB and $\Delta_{\text{SSIM}} = +0.01$, validating their importance for capturing global dependencies. Additionally, we analysed the impact of training dataset size on the MInDI-3D model using 64, 320, and 3200 subjects (Table 5). Increasing the training data size improved all metrics, with the 3200-subject model achieving the best metrics, i.e. $\Delta_{\text{MAE}} = -11.47$, $\Delta_{\text{PSNR}} = +3.72$ dB and $\Delta_{\text{SSIM}} = +0.03$ (compared to the 64-subject model), demonstrating the importance of training dataset size for deep-learning based artefact reduction in medical imaging.

We analyse the perception-distortion trade-off in MInDI-3D through progressive sampling (Figure 3). A single sampling step yields suboptimal results, failing to optimise either metric. Increasing steps beyond 2 (2-10 steps) trades distortion for realism: PSNR declines modestly (from 33.61 dB to 33.31 dB) while perceptual quality improves (FD DINOv2: from 75.83 to 20.14). This demonstrates that MInDI-3D enables controlled trade-offs between fidelity and realism through step adjustment. Visual examples of this trade-off for a lung tumour are shown in Figure 4, where added steps enhance sharpness and detail. The optimal amount of sampling steps for fidelity, varied across images and anatomic sites.

MInDI-3D achieves inference speeds competitive with the 3D U-Net baseline: a single sampling step requires 19 ms/volume versus the U-Nets 14 ms/volume (VRAM-loaded models). While MInDI-3D has a higher latency, its total runtime remains practical for clinical deployment, even at higher step counts (e.g., 10 steps require ≈190 ms for model inference).

3.2 Clinical Evaluation

To validate the quantitative results in a clinical setting, a MInDI-3D model – trained on sparse 50 volumes from 320 subjects – was tested on the real-

Dataset	MAE masked ↓	PSNR masked (dB) ↑	SSIM ↑
Uncorrected			
CT-RATE	134.04 ± 11.02	21.15 ± 0.69	0.29 ± 0.02
HyperSight	65.31 ± 8.56	27.45 ± 1.18	0.47 ± 0.01
MInDI-3D			
CT-RATE	20.70 ± 3.29	36.25 ± 1.24	0.97 ± 0.01
HyperSight	30.55 ± 4.44	33.61 ± 1.16	0.91 ± 0.01
3D U-Net			
CT-RATE	20.75 ± 3.50	36.18 ± 1.25	0.97 ± 0.01
HyperSight	32.16 ± 4.81	32.98 ± 1.18	0.90 ± 0.01

Table 2: Performance comparison of MInDI-3D (2-step inference) and 3D U-Net (equivalent architecture without time embedding) for correcting 50-projection reconstructions across CT-RATE (pseudo-CBCT) and HyperSight (real-world) datasets showing mean ± standard deviation. While both models achieve near-identical metrics on synthetic data (CT-RATE PSNR: 36.25 vs. 36.18, SSIM 0.97 vs 0.97), MInDI-3D performs slightly better than the U-Net on real-world HyperSight scans (PSNR 33.61 vs. 32.98, SSIM 0.91 vs. 0.90).

Projections	MAE masked ↓	PSNR masked (dB) ↑	SSIM ↑
Uncorrected			
25	125.29 ± 16.24	21.81 ± 1.15	0.32 ± 0.01
50	65.31 ± 8.56	27.45 ± 1.18	0.47 ± 0.01
100	27.84 ± 3.58	34.70 ± 1.21	0.70 ± 0.02
MInDI-3D			
25	48.03 ± 6.08	29.59 ± 1.00	0.86 ± 0.02
50	30.55 ± 4.44	33.61 ± 1.16	0.91 ± 0.01
100	24.62 ± 3.21	35.32 ± 0.94	0.93 ± 0.01

Table 3: MInDI-3Ds performance (2 step) across sparsity levels (25-100 projections) on the HyperSight dataset (MAE, PSNR, SSIM vs. ground truth (mean ± standard deviation)), where even 25-projection reconstructions achieve 62% lower MAE than uncorrected scans (48.02 vs. 125.29), validating its potential to enable ultra-low-dose CBCT.

Steps	DICE score				
	Lung Left	Lung Right	Vertebrae	Heart	Ribs
MInDI-3D					
1	0.96 ± 0.11	0.99 ± 0.00	0.95 ± 0.02	0.92 ± 0.03	0.89 ± 0.03
2	0.96 ± 0.11	0.99 ± 0.00	0.95 ± 0.02	0.92 ± 0.03	0.90 ± 0.03
3	0.96 ± 0.11	0.99 ± 0.00	0.95 ± 0.02	0.92 ± 0.03	0.90 ± 0.03
5	0.96 ± 0.11	0.99 ± 0.00	0.95 ± 0.02	0.92 ± 0.03	0.90 ± 0.03
10	0.96 ± 0.12	0.99 ± 0.00	0.95 ± 0.02	0.92 ± 0.03	0.90 ± 0.03
20	0.96 ± 0.12	0.99 ± 0.00	0.95 ± 0.02	0.91 ± 0.03	0.90 ± 0.03
30	0.96 ± 0.12	0.99 ± 0.00	0.95 ± 0.02	0.91 ± 0.03	0.90 ± 0.03
3D U-Net					
1	0.97 ± 0.09	0.99 ± 0.00	0.94 ± 0.02	0.92 ± 0.02	0.89 ± 0.03

Table 4: Stability of anatomical segmentation under iterative refinement (sparse 50). DICE scores (mean ± standard deviation) for MInDI-3D (1-30 sampling steps) vs. 3D U-Net on HyperSight CBCT data, benchmarked against full-dose ground-truth segmentations using Totalsegmentator. While U-Net achieves comparable performance in single-step inference, MInDI-3D maintains stable segmentation results across all anatomical structures (lung L/R: 0.96-0.99, vertebrae: 0.95, heart: 0.91-0.92, ribs: 0.89-0.90) despite 30 more sampling steps. Results exclude abdomen tumour patients because the analysed organs were not consistently present in their scans.

Configuration	MAE masked ↓	PSNR masked (dB) ↑	SSIM ↑
Uncorrected			
	134.04 ± 11.02	21.15 ± 0.69	0.29 ± 0.02
MInDI-3D: Dataset Size Ablation			
64 subjects	29.93 ± 7.15	33.53 ± 1.47	0.94 ± 0.03
320 subjects	21.10 ± 3.36	36.08 ± 1.23	0.96 ± 0.01
3200 subjects	18.46 ± 1.82	37.25 ± 0.84	0.97 ± 0.01
MInDI-3D: Attention Block Ablation			
no attention blocks	26.13 ± 7.92	34.06 ± 1.53	0.96 ± 0.02
1 attention blocks	22.09 ± 4.14	35.71 ± 1.40	0.97 ± 0.01
2 attention blocks	21.10 ± 3.36	36.08 ± 1.23	0.97 ± 0.01

Table 5: Ablation study of MInDI-3D (sparse 50, 1 step) performance on the CT-RATE dataset, evaluating (1) training data scalability (64-3200 subjects) and (2) attention block design (0-2 blocks). Larger datasets reduce reconstruction error (3200 subjects: MAE 18.46 vs. 29.93 for 64 subjects), while two attention blocks optimise long-range dependency modeling ($\Delta_{\text{PSNR}} = +2.02$ dB vs. no attention blocks). Metrics averaged over test volumes versus ground truth.

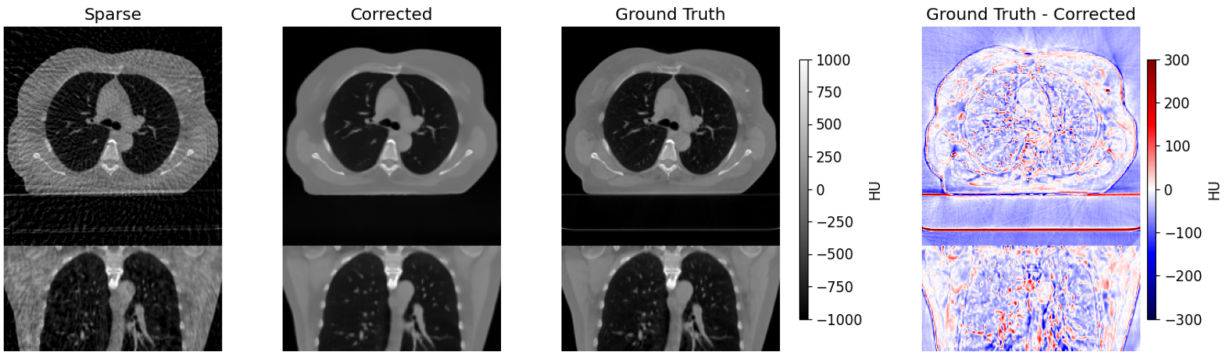

Figure 2: CBCT images (axial and coronal views) of a breast cancer patient (HyperSight dataset), from left to right showing the sparse volume (50 projections), corrected volume using the MInDI-3D model, ground truth volume and a difference plot (ground truth - corrected volume).

world HyperSight dataset. Performance was evaluated based on feedback from 11 clinicians from the Yonsei University Hospital, Seoul, South Korea. The real-world CBCT scans differed from the simulated training dataset, enabling assessment of the models' generalisation capabilities. The primary difference between the training and test datasets was the anatomic site: the training dataset consisted solely of chest CTs, while the test dataset included scans of the abdomen, breast, and lung. Additionally, the geometry used varied, with the training dataset employing half-fan and full-trajectory scans, and the test dataset using full-fan half-trajectory scans. We provided the clinicians with 16 paired CBCT volumes for review. The sparse volumes were corrected with the MInDI-3D model using 1 inference step and then set side-by side to the full-dose volumes. In every comparison, the tumour was highlighted on the planning CT for reference. The clinicians decided if the corrected sparse-view image was sufficient for any of the following tasks; positioning, contouring and/or dose calculation. The clinicians categorised themselves into the two general categories of radiation oncologist (64%) and medical physicist (36%).

For the task of patient positioning, a large part of clinicians agreed that this could be done using the enhanced CBCT volumes for all the anatomical sites investigated (abdomen 96.4%, lung & breast 100%). For the task of dose calculation and contouring, the responses were mixed. The acceptance rates for the AI-enhanced CBCT volumes for dose calculation were 40.0% for the abdomen, 54.6% for the breast and 69.7% for the lung scans. The acceptance rates for contouring were 41.8%, 80.0%, 90.9% for the anatomical sites abdomen, breast and lung respectively. Lung scans had the highest acceptance rate, while abdomen scans showed the lowest acceptance rate overall. Overall, the MInDI-3D model demonstrated strong clinical utility for patient positioning across all anatomical sites, with mixed but generally lower acceptance for dose calculation and contouring, particularly in the abdomen, highlighting a need for further refinement in these areas.

4 Discussion and Outlook

This work introduces MInDI-3D, the first, to our knowledge, adaptation of the InDI framework to 3D and adapted to the medical field. Our findings demonstrate that MInDI-3D not only effectively mitigates sparse-view artefacts, achieving quantitative performance comparable to a 3D U-Net, but also offers unique advantages in terms of a tuneable image quality.

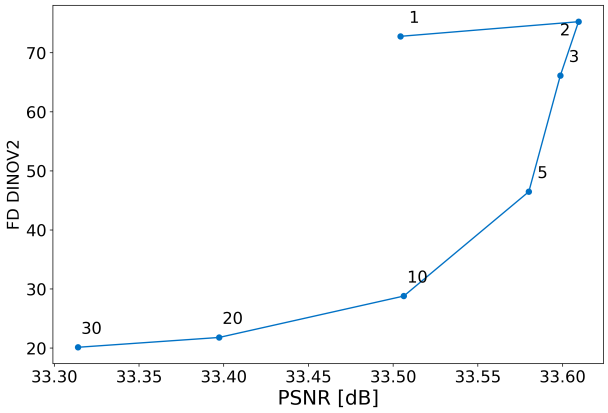

Figure 3: Perception-distortion trade-off in progressive sampling of MInDI-3D on the test set HyperSight with 50 projections. The lineplot compares fidelity (PSNR) and perceptual quality (FD DINOv2) across sampling steps (1-10). Sampling with 2-5 steps improves distortion (higher PSNR) compared to 1 step, while further steps enhance realism (lower FD DINOv2) at the expense of fidelity. Adjusting sampling steps enables precise control over realism and fidelity: steps beyond 2 prioritise perceptual quality, but optimal step counts may vary by anatomy.

We leverage a large-scale CT dataset via a pseudo-CBCT pipeline, and made the resulting dataset publicly available. This strategy successfully addresses the common limitation of data scarcity in medical imaging, and the observed scaling relationship between dataset size and performance (+3.72 dB PSNR gain) underscores its value. The model's robust generalisation across different anatomies, sparse-levels and unseen acquisition geometries is particularly encouraging. It suggests that MInDI-3D learns fundamental principles of artefact reduction rather than dataset-specific features.

The clinical relevance of MInDI-3D is multifaceted. Task-based evaluations confirm that its iterative refinements preserve crucial anatomical structures, maintaining high segmentation accuracy (e.g., lung DICE ≥ 0.96) even as perceptual quality is enhanced. This addresses a key concern with generative and deep learning models: ensuring that visual improvements do not compromise diagnostic or treatment-planning information. Direct clinical feedback supports the viability of MInDI-3D for clinical use in specific tasks, such as patient positioning (90-100%). The clinical tasks of dose calculation and contouring showed more variability between the anatomical sites. The superior acceptance rates for lung scans may reflect both inherent anatomical advantages (high contrast between tumour and surrounding tissue) and domain consistency between training and test data (chest).

While a direct comparison to other works is challenging due to differing reconstruction geometries, our results demonstrate competitive performance, as shown in Table 6. For example, Li et al. [21] reported 2D PSNR improvements of 15.56 and SSIM improvements of 0.636 for a sparse reconstruction with 29 projections. Similarly, Lee et al. [22] achieved 2D PSNR improvements of and SSIM of 0.951 with 36 projections. Li et al. [45] achieved a 2D PSNR of 31.29 and SSIM of 0.8471 with 30 projections. It's worth noting that these studies often include background (air) in their error calculations and used smaller datasets (only 10 volumes in [21,22]) for training and testing. Even with as few as 25 projections, we achieve a PSNR of 36.81 and an SSIM of 0.95 across the entire volume with improvements of 20.20 and 0.77 for PSNR and SSIM respectively.

The perception-distortion trade-off observed with MInDI-3D sampling steps mirrors findings in [19]. MInDI-3D users can adjust sampling steps to prioritise either quantitative fidelity or perceptual realism, tailoring the output to specific clinical needs. This flexibility is a key advantage, though finding the optimal balance and understanding its clinical implications across diverse scenarios remains an area for further investigation.

Three key considerations arise. First, the pseudo-CBCT simulation, while pragmatic, may not fully replicate real-world scatter and motion artefacts. Second, the perception metric (FD DINOv2) metric, though validated for natural images, requires clinical correlation with radiologist assessments in order to be validated for a clinical setting, building on [62]. Third, while diffusion-based models risk introducing

Method	Projections	PSNR	SSIM	PSNR Improvement	SSIM Improvement
Related Work					
Lee et al. [22]	36	38.25 (2D)	0.949* (2D)	–	–
Li et al. [21]	29	38.21* (2D)	0.936* (2D)	+15.56*	+0.636*
Li et al. [45]	30	31.29 (2D)	0.847 (2D)	+12.98	+0.617
Our Work: MInDI-3D					
Validation set (CT-RATE)	50	**41.63** (3D)	**0.97** (3D)	**+20.48**	+0.68
Test set (HyperSight)	50	30.34 (3D)	0.91 (3D)	+ 6.36	+0.44
Validation set (CT-RATE)	25	36.81 (3D)	0.95 (3D)	+20.20	**+0.77**
Test set (HyperSight)	25	29.30 (3D)	0.86 (3D)	+10.00	+0.54
Our Work: MInDI-3D (body masked)					
Validation set (CT-RATE)	50	37.29 (3D, masked)	0.97 (3D)	+12.96	+0.68
Test set (HyperSight)	50	33.55 (3D, masked)	0.91 (3D)	+ 6.10	+0.44
Validation set (CT-RATE)	25	32.50 (3D, masked)	0.95 (3D)	+13.45	**+0.77**
Test set (HyperSight)	25	29.59 (3D, masked)	0.86 (3D)	+ 7.78	+0.54

Table 6: We compare our best results in terms of reconstruction quality with related work. Improvements are computed relative to the analytical FBP reconstruction. * Indicates an average over the three planes while **bold** indicates the best results.

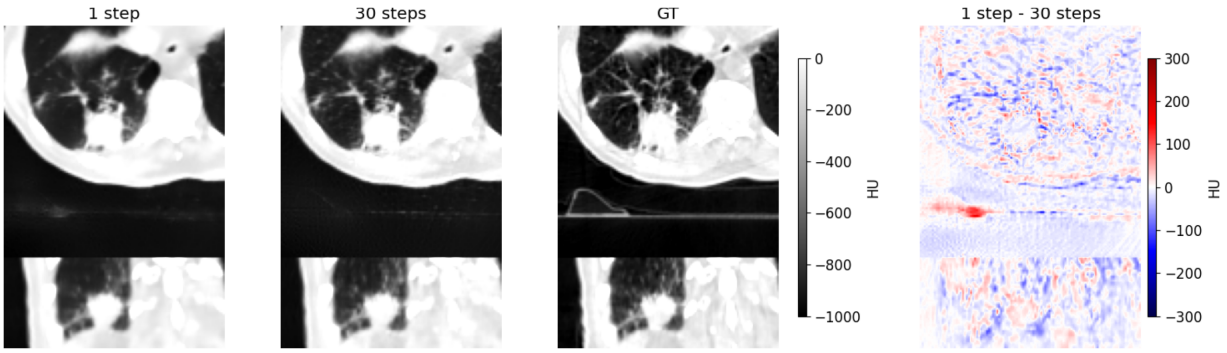

Figure 4: Comparing the MInDI-3D prediction of a lung tumour (lower right lung lobe) from a sparse 50 reconstruction with 1 vs. 30 steps (the ground truth and the difference of step 1 - step 30 as reference). There is an increase of sharpness and detail from step 1 to step 30

synthetic anatomical features that could mislead clinical interpretation, a critical concern in safety-critical applications like radiotherapy planning, we proactively mitigated this risk through a clinical evaluation. However, future work should rigorously test whether increased sampling steps (3 sampling steps and above) retain anatomical accuracy. This could be done by calculating the treatment dose at different sampling steps to determine whether the dose estimation remains consistent.

Future work should further investigate the trade-off between perceived image quality and anatomical fidelity. Specifically, it is necessary to determine if adding more iteration steps improves clinical usability or if it inadvertently diminishes anatomical accuracy or enhances remaining artefacts. In this context a systematic study could be conducted on how perception metrics (e.g., FD DINOv2) that were trained on natural images can be utilised to measure perception in 3D in a medical setting. While this study has focused on image enhancement in the pixel space further research could be conducted in enhancing images in a latent space, which should allow to train deep learning models with a higher resolution in 3D.

Our implementation of MInDI-3D establishes conditional generative-based models as viable tools for sparse-view CBCT restoration, achieving clinically acceptable image quality across multiple anatomical sites for certain tasks related to radiation therapy. The framework's generalisation across datasets and scaling with training size highlights the potential of large-scale 3D medical imaging models to advance adaptive radiotherapy.

Conflict of Interest

The authors declare no conflicts of interest related to this work. Some authors are employed by Varian, as indicated in the affiliations. Their involvement in the study was in accordance with standard academic and ethical practices.

References

[1] D. A. Jaffray, J. H. Siewerdsen, J. W. Wong, and A. A. Martinez, "Flat-panel cone-beam computed tomography for image-guided radiation therapy," *International Journal of Radiation Oncology, Biology, Physics*, vol. 53, no. 5, pp. 1337–1349, Aug. 2002, publisher: Elsevier. [Online]. Available: https://www.redjournal.org/article/S0360-3016(02)02884-5/abstract

[2] S. W. Yoon, H. Lin, M. Alonso-Basanta, N. Anderson, O. Apinorasethkul, K. Cooper,

L. Dong, B. Kempsey, J. Marcel, J. Metz, R. Scheuermann, and T. Li, "Initial Evaluation of a Novel Cone-Beam CT-Based Semi-Automated Online Adaptive Radiotherapy System for Head and Neck Cancer Treatment A Timing and Automation Quality Study," *Cureus*, Aug. 2020, publisher: Springer Science and Business Media LLC. [Online]. Available: https://www.cureus.com/articles/35411-initial-evaluation-of-a-novel-cone-beam-ct-based-semi-automated-online-adaptive-radiotherapy-system-for-head-

[3] E. Lavrova, M. D. Garrett, Y.-F. Wang, C. Chin, C. Elliston, M. Savacool, M. Price, L. A. Kachnic, and D. P. Horowitz, "Adaptive Radiation Therapy: A Review of CT-based Techniques," *Radiology: Imaging Cancer*, vol. 5, no. 4, p. e230011, Jul. 2023, publisher: Radiological Society of North America. [Online]. Available: https://pubs.rsna.org/doi/full/10.1148/rycan.230011

[4] M. Amirian, D. Barco, I. Herzig, and F.-P. Schilling, "Artifact Reduction in 3D and 4D Cone-beam Computed Tomography Images with Deep Learning - A Review," *IEEE Access*, pp. 1–1, 2024, conference Name: IEEE Access. [Online]. Available: https://ieeexplore.ieee.org/document/10398205

[5] S. Sharma and K. Guleria, "Deep Learning Models for Image Classification: Comparison and Applications," in *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, Apr. 2022, pp. 1733–1738. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9823516

[6] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image Segmentation Using Deep Learning: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3523–3542, Jul. 2022, conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9356353

[7] L. Zhai, Y. Wang, S. Cui, and Y. Zhou, "A Comprehensive Review of Deep Learning-Based Real-World Image Restoration," *IEEE Access*, vol. 11, pp. 21 049–21 067, 2023, conference Name: IEEE Access. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10056934

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham: Springer International Publishing, 2015, pp. 234–241. [Online]. Available: https://doi.org/10.1007/978-3-319-24574-4_28

[9] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," in *Advances in Neural Information Processing Systems*, vol. 27. Curran Associates, Inc., 2014. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2014/hash/f033ed80deb0234979a61f95710dbe25-Abstract.html

[10] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," Dec. 2013. [Online]. Available: https://arxiv.org/abs/1312.6114v11

[11] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 6840–6851. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf

[12] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis With Latent Diffusion Models," 2022, pp. 10 684–10 695. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2022/html/Rombach_

High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.html

[13] A. Mumuni, F. Mumuni, and N. K. Gerrar, "A Survey of Synthetic Data Augmentation Methods in Machine Vision," *Machine Intelligence Research*, vol. 21, no. 5, pp. 831–869, Oct. 2024. [Online]. Available: https://doi.org/10.1007/s11633-022-1411-7

[14] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-To-Image Translation With Conditional Adversarial Networks," 2017, pp. 1125–1134. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2017/html/Isola_Image-To-Image_Translation_With_CVPR_2017_paper.html

[15] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 2242–2251, iSSN: 2380-7504. [Online]. Available: https://ieeexplore.ieee.org/document/8237506

[16] M. Ali, M. Ali, M. Hussain, and D. Koundal, "Generative Adversarial Networks (GANs) for Medical Image Processing: Recent Advancements," *Archives of Computational Methods in Engineering*, Oct. 2024. [Online]. Available: https://doi.org/10.1007/s11831-024-10174-8

[17] P. Dhariwal and A. Nichol, "Diffusion Models Beat GANs on Image Synthesis," in *Advances in Neural Information Processing Systems*, vol. 34. Curran Associates, Inc., 2021, pp. 8780–8794. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/hash/49ad23d1ec9fa4bd8d77d02681df5cfa-Abstract.html

[18] G. Mller-Franzes, L. Huck, M. Bode, S. Nebelung, C. Kuhl, and D. Truhn, "Diffusion probabilistic versus generative adversarial models to reduce contrast agent dose in breast MRI," *European Radiology Experimental*, vol. 8, no. 1, p. 53, May 2024. [Online]. Available: https://doi.org/10.1186/s41747-024-00451-3

[19] M. Delbracio and P. Milanfar, "Inversion by Direct Iteration: An Alternative to Denoising Diffusion for Image Restoration," *Transactions on Machine Learning Research*, Mar. 2023. [Online]. Available: https://openreview.net/forum?id=VmyFF5lL3F

[20] J. He, B. Li, G. Yang, and Z. Liu, "Blaze3DM: Marry Triplane Representation with Diffusion for 3D Medical Inverse Problem Solving," May 2024. [Online]. Available: http://arxiv.org/abs/2405.15241

[21] Z. Li, Y. Wang, J. Zhang, W. Wu, and H. Yu, "Two-and-a-half Order Score-based Model for Solving 3D Ill-posed Inverse Problems," *Computers in Biology and Medicine*, vol. 168, p. 107819, Jan. 2024. [Online]. Available: http://arxiv.org/abs/2308.08511

[22] S. Lee, H. Chung, M. Park, J. Park, W.-S. Ryu, and J. C. Ye, "Improving 3D Imaging with Pre-Trained Perpendicular 2D Diffusion Models," 2023, pp. 10 710–10 720. [Online]. Available: https://openaccess.thecvf.com/content/ICCV2023/html/Lee_Improving_3D_Imaging_with_Pre-Trained_Perpendicular_2D_Diffusion_Models_ICCV_2023_paper.html

[23] H. Chung, D. Ryu, M. T. Mccann, M. L. Klasky, and J. C. Ye, "Solving 3D Inverse Problems Using Pre-Trained 2D Diffusion Models," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Vancouver, BC, Canada: IEEE, Jun. 2023, pp. 22 542–22 551. [Online]. Available: https://ieeexplore.ieee.org/document/10204965/

[24] Y. Blau and T. Michaeli, "The Perception-Distortion Tradeoff," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 6228–6237, arXiv:1711.06077 [cs]. [Online]. Available: http://arxiv.org/abs/1711.06077

[25] R. Schulze, U. Heil, D. Gross, D. Bruellmann, E. Dranischnikow, U. Schwanecke, and E. Schoemer, "Artefacts in CBCT: a review," *Dentomaxillofacial Radiology*, vol. 40, no. 5, pp. 265–273, Jul. 2011. [Online]. Available: https://doi.org/10.1259/dmfr/30642039

[26] F. E. Boas and D. Fleischmann, "CT artifacts: causes and reduction techniques," *Imaging in Medicine*, vol. 4, no. 2, pp. 229–240, Apr. 2012, publisher: Open Access Journals. [Online]. Available: https://www.openaccessjournals.com/abstract/ct-artifacts-causes-and-reduction-techniques-9353.html

[27] M. Amirian, J. A. Montoya-Zegarra, I. Herzig, P. Eggenberger Hotz, L. Lichtensteiger, M. Morf, A. Zst, P. Paysan, I. Peterlik, S. Scheib, R. M. Fchslin, T. Stadelmann, and F.-P. Schilling, "Mitigation of motion-induced artifacts in cone beam computed tomography using deep convolutional neural networks," *Medical Physics*, vol. 50, no. 10, pp. 6228–6242, 2023. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/mp.16405

[28] Y. Wang, L. Chao, W. Shan, H. Zhang, Z. Wang, and Q. Li, "Improving the Quality of Sparse-view Cone-Beam Computed Tomography via Reconstruction-Friendly Interpolation Network," in *Computer Vision ACCV 2022*, L. Wang, J. Gall, T.-J. Chin, I. Sato, and R. Chellappa, Eds. Cham: Springer Nature Switzerland, 2023, pp. 86–100. [Online]. Available: https://doi.org/10.1007/978-3-031-26351-4_6

[29] D. Hu, Y. Zhang, J. Liu, Y. Zhang, J. L. Coatrieux, and Y. Chen, "PRIOR: Prior-Regularized Iterative Optimization Reconstruction For 4D CBCT," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 11, pp. 5551–5562, Nov. 2022, conference Name: IEEE Journal of Biomedical and Health Informatics. [Online]. Available: https://ieeexplore.ieee.org/document/9866113

[30] Z. Jiang, Z. Zhang, Y. Chang, Y. Ge, F.-F. Yin, and L. Ren, "Prior image-guided cone-beam computed tomography augmentation from under-sampled projections using a convolutional neural network," *Quantitative Imaging in Medicine and Surgery*, vol. 11, no. 12, pp. 4 767 780–4 764 780, Dec. 2021, publisher: AME Publishing Company. [Online]. Available: https://qims.amegroups.org/article/view/75305

[31] F. Khader, G. Mller-Franzes, S. Tayebi Arasteh, T. Han, C. Haarburger, M. Schulze-Hagen, P. Schad, S. Engelhardt, B. Baeler, S. Foersch, J. Stegmaier, C. Kuhl, S. Nebelung, J. N. Kather, and D. Truhn, "Denoising diffusion probabilistic models for 3D medical image generation," *Scientific Reports*, vol. 13, no. 1, p. 7303, May 2023, number: 1 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41598-023-34341-2

[32] S. Kim, D. C. Alexander, A. K. Eldaly, M. Figini, and H. F. J. Tregidgo, "A 3D Conditional Diffusion Model for Image Quality Transfer - An Application to Low-Field MRI," Oct. 2023. [Online]. Available: https://openreview.net/forum?id=TynSiNAVc8

[33] Y. Liu, G. Dwivedi, F. Boussaid, F. Sanfilippo, M. Yamada, and M. Bennamoun, "Inflating 2D convolution weights for efficient generation of 3D medical images," *Computer Methods and Programs in Biomedicine*, vol. 240, p. 107685, Oct. 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169260723003504

[34] A. Volokitin, E. Erdil, N. Karani, K. C. Tezcan, X. Chen, L. Van Gool, and E. Konukoglu, "Modelling the Distribution of 3D Brain MRI Using a 2D Slice VAE," in *Medical Image Computing and Computer Assisted Intervention MICCAI 2020*, A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu, and L. Joskowicz, Eds. Cham: Springer International Publishing, 2020, pp. 657–666. [Online]. Available: https://doi.org/10.1007/978-3-030-59728-3_64

[35] J. Kapoor, J. H. Macke, and C. F. Baumgartner, "Multiscale Metamorphic VAE for 3D Brain MRI Synthesis," Jan. 2023, arXiv:2301.03588 [eess]. [Online]. Available: http://arxiv.org/abs/2301.03588

[36] P. Friedrich, J. Wolleb, F. Bieder, A. Durrer, and P. C. Cattin, "WDM: 3D Wavelet Diffusion Models forHigh-Resolution Medical Image Synthesis," in *Deep Generative Models*, A. Mukhopadhyay, I. Oksuz, S. Engelhardt, D. Mehrof, and Y. Yuan, Eds. Cham: Springer Nature Switzerland, 2025, pp. 11–21. [Online]. Available: https://doi.org/10.1007/978-3-031-72744-3_2

[37] S. Pan, E. Abouei, J. Wynne, C.-W. Chang, T. Wang, R. L. J. Qiu, Y. Li, J. Peng, J. Roper, P. Patel, D. S. Yu, H. Mao, and X. Yang, "Synthetic CT generation from MRI using 3D transformer-based denoising diffusion model," *Medical Physics*, vol. 51, no. 4, pp. 2538–2548, 2024. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/mp.16847

[38] Z. Dorjsembe, H.-K. Pao, S. Odonchimed, and F. Xiao, "Conditional Diffusion Models for Semantic 3D Brain MRI Synthesis," *IEEE Journal of Biomedical and Health Informatics*, 2024, publisher: IEEE. [Online]. Available: https://doi.org/10.1109/jbhi.2024.3385504

[39] Y. Wang, Y. Luo, C. Zu, B. Zhan, Z. Jiao, X. Wu, J. Zhou, D. Shen, and L. Zhou, "3D multi-modality Transformer-GAN for high-quality PET reconstruction," *Medical Image Analysis*, vol. 91, p. 102983, Jan. 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841523002438

[40] S. Poonkodi and M. Kanchana, "3d-medtrancsgan: 3d medical image transformation using csgan," *Computers in Biology and Medicine*, vol. 153, p. 106541, 2023, publisher: Elsevier. [Online]. Available: https://doi.org/10.1016/j.compbiomed.2023.106541

[41] P. Friedrich, Y. Frisch, and P. C. Cattin, "Deep Generative Models for 3D Medical Image Synthesis," in *Generative Machine Learning Models in Medical Image Computing*, L. Zhang, C. Chen, Z. Li, and G. Slabaugh, Eds. Cham: Springer Nature Switzerland, 2025, pp. 255–278. [Online]. Available: https://doi.org/10.1007/978-3-031-80965-1_13

[42] Y. Xue, Y. Peng, L. Bi, D. Feng, and J. Kim, "CG-3DSRGAN: A classification guided 3D generative adversarial network for image quality recovery from low-dose PET images," in *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, Jul. 2023, pp. 1–4, iSSN: 2694-0604. [Online]. Available: https://ieeexplore.ieee.org/document/10341112

[43] P. Zeng, L. Zhou, C. Zu, X. Zeng, Z. Jiao, X. Wu, J. Zhou, D. Shen, and Y. Wang, "3D CVT-GAN: A 3D Convolutional Vision Transformer-GAN for PET Reconstruction," in *Medical Image Computing and Computer Assisted Intervention MICCAI 2022*, L. Wang, Q. Dou, P. T. Fletcher, S. Speidel, and S. Li, Eds. Cham: Springer Nature Switzerland, 2022, pp. 516–526. [Online]. Available: https://doi.org/10.1007/978-3-031-16446-0_49

[44] C. McCollough, "TU-FG-207A-04: Overview of the Low Dose CT Grand Challenge," *Medical Physics*, vol. 43, no. 6Part35, pp. 3759–3760, 2016. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1118/1.4957556

[45] S. Li, X. Jiang, M. Tivnan, G. J. Gang, Y. Shen, and J. W. Stayman, "CT Reconstruction using Diffusion Posterior Sampling conditioned on a Nonlinear Measurement Model," *Journal of Medical Imaging*, vol. 11, no. 04, Aug. 2024, arXiv:2312.01464 [physics]. [Online]. Available: http://arxiv.org/abs/2312.01464

[46] H. R. Roth, L. Lu, A. Seff, K. M. Cherry, J. Hoffman, S. Wang, J. Liu, E. Turkbey, and R. M. Summers, "A new 2.5D representation for lymph node detection using random sets of deep

convolutional neural network observations," *Medical image computing and computer-assisted intervention: MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention*, vol. 17, no. Pt 1, pp. 520–527, 2014. [Online]. Available: https://doi.org/10.1007/978-3-319-10404-1_65

[47] I. E. Hamamci, S. Er, F. Almas, A. G. Simsek, S. N. Esirgun, I. Dogan, M. F. Dasdelen, B. Wittmann, E. Simsar, M. Simsar, E. B. Erdemir, A. Alanbay, A. Sekuboyina, B. Lafci, M. K. zdemir, and B. H. Menze, "A foundation model utilizing chest CT volumes and radiology reports for supervised-level zero-shot detection of abnormalities," *CoRR*, Jan. 2024. [Online]. Available: https://openreview.net/forum?id=kurajHzp19&trk=public_post_comment-text

[48] L. A. Feldkamp, L. C. Davis, and J. W. Kress, "Practical cone-beam algorithm," *JOSA A*, vol. 1, no. 6, pp. 612–619, Jun. 1984, publisher: Optica Publishing Group. [Online]. Available: https://opg.optica.org/josaa/abstract.cfm?uri=josaa-1-6-612

[49] A. H. Andersen and A. C. Kak, "Simultaneous Algebraic Reconstruction Technique (SART): A superior implementation of the ART algorithm," *Ultrasonic Imaging*, vol. 6, no. 1, pp. 81–94, Jan. 1984. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0161734684900087

[50] G. N. Ramachandran and A. V. Lakshminarayanan, "Three-dimensional Reconstruction from Radiographs and Electron Micrographs: Application of Convolutions instead of Fourier Transforms," *Proceedings of the National Academy of Sciences*, vol. 68, no. 9, pp. 2236–2240, Sep. 1971, publisher: Proceedings of the National Academy of Sciences. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.68.9.2236

[51] . iek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation," in *Medical Image Computing and Computer-Assisted Intervention MICCAI 2016*, S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, and W. Wells, Eds. Cham: Springer International Publishing, 2016, pp. 424–432. [Online]. Available: https://doi.org/10.1007/978-3-319-46723-8_49

[52] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," in *Proceedings of the 32nd International Conference on Machine Learning*. PMLR, Jun. 2015, pp. 448–456, iSSN: 1938-7228. [Online]. Available: https://proceedings.mlr.press/v37/ioffe15.html

[53] D. Hendrycks and K. Gimpel, "Gaussian Error Linear Units (GELUs)," Jun. 2016. [Online]. Available: https://arxiv.org/abs/1606.08415v5

[54] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, . u. Kaiser, and I. Polosukhin, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[55] P. Esser, R. Rombach, and B. Ommer, "Taming Transformers for High-Resolution Image Synthesis," 2021, pp. 12 873–12 883. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2021/html/Esser_Taming_Transformers_for_High-Resolution_Image_Synthesis_CVPR_2021_paper.html?ref=

[56] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," Jan. 2017, arXiv:1412.6980 [cs]. [Online]. Available: http://arxiv.org/abs/1412.6980

[57] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, Apr. 2004, conference Name: IEEE Transactions

on Image Processing. [Online]. Available: https://ieeexplore.ieee.org/document/1284395

[58] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, Jan. 1979. [Online]. Available: https://ieeexplore.ieee.org/document/4310076

[59] G. Stein, J. Cresswell, R. Hosseinzadeh, Y. Sui, B. Ross, V. Villecroze, Z. Liu, A. L. Caterini, E. Taylor, and G. Loaiza-Ganem, "Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models," *Advances in Neural Information Processing Systems*, vol. 36, pp. 3732–3784, Dec. 2023. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/hash/0bc795afae289ed465a65a3b4b1f4eb7-Abstract-Conference.html

[60] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "DINOv2: Learning Robust Visual Features without Supervision," *Transactions on Machine Learning Research Journal*, p. 1, Jan. 2024. [Online]. Available: https://hal.science/hal-04376640

[61] J. Wasserthal, H.-C. Breit, M. T. Meyer, M. Pradella, D. Hinck, A. W. Sauter, T. Heye, D. T. Boll, J. Cyriac, S. Yang, M. Bach, and M. Segeroth, "TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images," *Radiology: Artificial Intelligence*, vol. 5, no. 5, p. e230024, Sep. 2023, publisher: Radiological Society of North America. [Online]. Available: https://pubs.rsna.org/doi/10.1148/ryai.230024

[62] M. Woodland, A. Castelo, M. Al Taie, J. Albuquerque Marques Silva, M. Eltaher, F. Mohn, A. Shieh, S. Kundu, J. P. Yung, A. B. Patel, and K. K. Brock, "Feature Extraction forGenerative Medical Imaging Evaluation: New Evidence Against anEvolving Trend," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, M. G. Linguraru, Q. Dou, A. Feragen, S. Giannarou, B. Glocker, K. Lekadir, and J. A. Schnabel, Eds. Cham: Springer Nature Switzerland, 2024, pp. 87–97. [Online]. Available: https://doi.org/10.1007/978-3-031-72390-2_9