IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 17 December 2025, accepted 31 December 2025, date of publication 12 January 2026, date of current version 15 January 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3652627

RESEARCH ARTICLE

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), 8401 Winterthur, Switzerland
[2]Institute of Applied Mathematics and Physics (IAMP), Zurich University of Applied Sciences (ZHAW), 8401 Winterthur, Switzerland
[3]Varian Medical Systems Imaging Laboratory, 5405 Baden, Switzerland
[4]Department of Quantitative Biomedicine, University of Zurich, 8006 Zurich, Switzerland

Corresponding author: Daniel Barco (baoc@zhaw.ch)

This work was supported by Innosuisse under Grant 120.480 IP-LS.

ABSTRACT Reducing patient radiation exposure in Cone Beam Computed Tomography (CBCT) by acquiring fewer projections introduces severe image artefacts, limiting its clinical utility. To address this challenge, we propose **MInDI-3D** (**M**edical **In**version by **D**irect **I**teration in **3D**), a 3D conditional diffusion framework that restores volumetric data from sparse-view inputs. Our work provides two key contributions: 1) The MInDI-3D model, the first adaptation of the iterative inversion principle to fully 3D medical volumes, which offers a unique, tuneable trade-off between perceptual quality and quantitative fidelity by adjusting the number of inference steps. 2) A new, publicly available, large-scale dataset of 16,182 pseudo-CBCT volumes to facilitate robust training and future research. On an independent real-world CBCT test set, MInDI-3D achieves performance competitive with state-of-the-art methods, yielding a 0.54 SSIM gain over standard reconstructions from only 25 projections. This result enables a 16-fold reduction in radiation exposure and demonstrates robust generalisation to a new scanner geometry not seen during training. Beyond standard metrics, MInDI-3D reconstructions preserved high anatomical integrity, enabling accurate automated segmentation in task-based evaluations. In a clinical evaluation by 11 radiotherapy specialists, the reconstructions were rated as sufficient for patient positioning across all tested anatomical sites (abdomen, breast, and lung) and were noted to preserve lung tumour boundaries well.

INDEX TERMS Cone beam computed tomography (CBCT), deep learning, sparse-view artefacts.

I. 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) exemplifies 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

The associate editor coordinating the review of this manuscript and approving it for publication was Kumaradevan Punithakumar.

treatment periods (up to 40 sessions) raise concerns about cumulative radiation exposure to 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 Nyquist–Shannon 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 models, particularly convolutional neural networks (CNNs) like the U-Net [5], are well-suited for medical image reconstruction [6]. The U-Net's encoder-decoder architecture has inspired many variants, including the 3D U-Net for volumetric CBCT data. More recently, generative models such as GANs [7], VAEs [8], and diffusion models [9] have been extended to conditional image-to-image tasks relevant to medical imaging, including artefact removal, domain translation, and denoising [10].

Diffusion models have significantly advanced image synthesis and restoration, surpassing traditional GANs in conditional and unconditional generation tasks [11], [12], [13]. 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 [9], [12]. This makes them impractical for many real-world applications, due to prolonged inference times. To address this, InDI (Inversion by Direct Iteration) [14] was proposed as an efficient alternative for conditional 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 [15], [16], [17], [18]. 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 artefact removal.

Our study compares the performance of MInDI-3D with state-of-the-art models. We evaluate the impact of varying training dataset 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) [19]. InDI enables control over this trade-off without retraining: increasing the number of sampling steps, InDI can trade distortion for better perception reducing the problem of regression to the mean by adding realistic features [14]. We explore this perception-distortion trade-off.

Our main contributions are summarised as follows:

- We introduce MInDI-3D, the first iterative diffusion model adapted for fully volumetric sparse-view CBCT, which provides a unique trade-off between quantitative fidelity and perceptual quality.
- We present and publicly release a large-scale pseudo-CBCT dataset of 16,182 volumes to address data scarcity and provide a benchmark for future research in 3D medical image restoration.
- We provide a comprehensive validation demonstrating that MInDI-3D enables radiation reduction while maintaining clinically sufficient image quality, confirmed through rigorous quantitative, generalisation, and task-based clinical evaluations by 11 clinicians.

A. RELATED WORK

Extensive research has been conducted on characterising and mitigating artefacts that degrade image quality in CT and CBCT reconstruction [20], [21]. 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. [22] 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 [23], [24], [25]. 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 privacy-preserving tool to augment small medical image datasets [26]. Three main architecture types have been used in 3D unconditional medical image generation: GANs [27], [28], VAEs [29], [30] and diffusion models [26], [31]. 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 [32], [33], [34], [35]. 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 [36]. 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 [34], [37], [38], recent efforts have shifted toward diffusion-based approaches.

Our work is most closely related to diffusion-based models used for sparse-view reconstruction, which often leverage multiple 2D diffusion models. For instance, DiffusionMBIR [18] proposes an effective method for 3D reconstruction by augmenting a pre-trained 2D diffusion model prior with a model-based Total Variation (TV) prior in the z-direction. This TV prior enforces coherence between slices, and the entire process is integrated into an iterative reconstruction framework that includes measurement consistency. Building on the concept of using a 2D diffusion model, the Two-and-a-half-D Score Matching (TOSM) model [16] utilizes a 2.5D fusion technique that combines scores from three orthogonal planes (sagittal, coronal, and transversal) derived from a single pre-trained 2D model to approximate a full 3D score. Similarly, Two Perpendicular 2D Diffusion Models (TPDM) [17] models the 3D data distribution as a product of two perpendicular 2D plane distributions, performing posterior-based sampling alternatively in each direction. Blaze3DM [15] integrates a triplane neural field representation with a diffusion model. This approach first constructs compact, data-dependent triplane embeddings from the 3D volumes and then trains a diffusion model on the distribution of these efficient embeddings, significantly reducing computational load. DiffusionMBIR, TOSM, TPDM and Blaze3DM were evaluated on the AAPM Low Dose CT Grand Challenge dataset [39], using 9 pseudo CBCT volumes for training and 1 for testing. In contrast, Diffusion Posterior Sampling (DPS) [40] trained on the larger CT Lymph Nodes Dataset [41] while comparing different diffusion-based methods.

These varied strategies highlight a broader trend towards computationally efficient pseudo-3D solutions [36]. The computational burdens of diffusion-based models have motivated research into efficient diffusion-based implementations that can be practically applied in a clinical setting where reconstruction speed is essential. In this context, InDI is particularly promising, as it requires only a fraction of the sampling steps compared to other diffusion-based models for the conditional setting [14].

II. MATERIALS & METHODS

This section outlines the proposed MInDI-3D framework. We first introduce the data generation pipeline used to address data scarcity, followed by a detailed description of the network architecture. Finally, we present the adaptation of the iterative inversion process to fully 3D volumetric data and the training protocol employed.

A. DATASETS

CT-RATE is a public dataset [42] 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 × 0.72 × 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 II-B), which can then be reconstructed by a CBCT reconstruction algorithm to mimic the CBCT acquisition. This dataset (3,612 patients) is orders of magnitude larger than typical public benchmarks used in prior diffusion-based CT reconstruction studies, such as the AAPM Low Dose CT Grand Challenge dataset (approx. 10 patients), thereby enabling more robust model generalization. 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 Hugging Face.[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.

B. CBCT RECONSTRUCTION AND SIMULATION

Reconstructing 3D CBCT volumes from 2D projections can be achieved through analytical and iterative approaches. The Feldkamp-Davis-Kress (FDK) [43] algorithm, an analytical method, provides a fast and reliable approximation of the inverse Radon transform, establishing itself as a widely used baseline for 3D CBCT reconstruction. While FDK excels in computational efficiency, iterative reconstruction techniques – such as the Simultaneous Algebraic Reconstruction Technique (SART) [44] – 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 [45] 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 overlapping

[1] https://huggingface.co/datasets/danielbarco/sparse-CT-RATE

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.

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

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 × 256 × 64 voxels and a spacing of 2 × 2 × 3 mm. We chose this volume size and spacing to balance memory constraints in our 3D deep learning pipelines with anatomical coverage.

C. 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 [5], [46] (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.

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 × 3 × 3, stride: 1). The residual submodules process the volume as follows: (1) batch normalisation [47] (BN), (2) SiLU activation [48], and (3) a 3D convolution (kernel: 3 × 3 × 3, stride: 1). The input to the residual submodule is then added to the output. After the residual blocks, a strided convolution (kernel: 3 × 3 × 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: 4×4×4, stride: 2). Finally, on the last stage, an additional 3D convolutional layer (kernel: 3×3×3, stride: 1) is employed. Channel dimensions halve at each stage, reversing the encoder's progression (512 to 32).

Attention Mechanism: Convolutional attention applies the Scaled Dot-Product Attention [49] to a convolutional layer following [50]. The convolutional attention mechanism is integrated into the deepest two encoder and decoder layers and is described subsequently. Input features first undergo group normalisation, followed by a 1 × 1 × 1 convolution 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) 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 [14]. 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

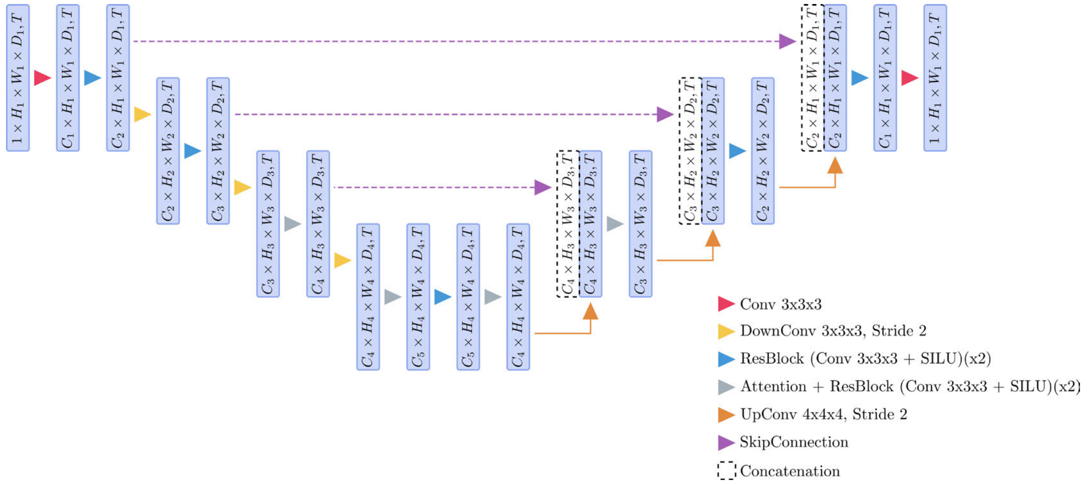

$1 \times H_1 \times W_1 \times D_1, T$
$C_1 \times H_1 \times W_1 \times D_1, T$
$C_2 \times H_1 \times W_1 \times D_1, T$
$C_2 \times H_2 \times W_2 \times D_2, T$
$C_3 \times H_2 \times W_2 \times D_2, T$
$C_3 \times H_3 \times W_3 \times D_3, T$
$C_4 \times H_3 \times W_3 \times D_3, T$
$C_4 \times H_4 \times W_4 \times D_4, T$
$C_5 \times H_4 \times W_4 \times D_4, T$
$C_5 \times H_4 \times W_4 \times D_4, T$
$C_4 \times H_4 \times W_4 \times D_4, T$
$C_4 \times H_3 \times W_3 \times D_3, T$
$C_4 \times H_3 \times W_3 \times D_3, T$
$C_3 \times H_3 \times W_3 \times D_3, T$
$C_3 \times H_2 \times W_2 \times D_2, T$
$C_3 \times H_2 \times W_2 \times D_2, T$
$C_2 \times H_2 \times W_2 \times D_2, T$
$C_2 \times H_1 \times W_1 \times D_1, T$
$C_2 \times H_1 \times W_1 \times D_1, T$
$C_1 \times H_1 \times W_1 \times D_1, T$
$1 \times H_1 \times W_1 \times D_1, T$
Conv 3x3x3
DownConv 3x3x3, Stride 2
ResBlock (Conv 3x3x3 + SILU)(x2)
Attention + ResBlock (Conv 3x3x3 + SILU)(x2)
UpConv 4x4x4, Stride 2
SkipConnection
Concatenation

FIGURE 1. U-Net architecture with 4 hierarchical levels, showing layer-specific dimensionality (*C* × *H* × *W* × *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.

$t = 0$.

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

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

We incorporate a time-embedding into the U-Net backbone of the MInDI-3D 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 [9] with 1024 channels.

D. TRAINING

Training is conducted on an NVIDIA H200 GPU with 140 GB of VRAM, using the Adam optimiser [51] (learning rate 0.0001) and mean absolute error (MAE) (cf. II-E) 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 optimisation was conducted via a combination of Bayesian Optimisation over common parameter ranges and manual tuning, with the final configuration selected based on the best-performing model on the validation set, as measured by the Mean Absolute Error (MAE) loss. 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. The high memory footprint was a deliberate choice to enable training on full 3D volumes ($256 \times 256 \times 64$) directly in the voxel space, avoiding patching or compression which could compromise image quality.

E. 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 $\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|x_i - x_i'|$, 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) [52];
- Peak Signal-to-Noise Ratio $\text{PSNR} = 10\log_{10}\left(\frac{\text{MAX}^2}{\frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2}\right)$, 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, and Table 4, 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 [53] 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 Fréchet Distance (FD) [54] 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 [55], which as feature extractor has shown to most closely align with human perception and is superior to FID based on Inception [54]. 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 [56] to segment the heart, left lung, right lung, ribs, and vertebrae. We chose TotalSegmentator as a segmentation tool due to its robust segmentation capabilities.

III. RESULTS

A. QUANTITATIVE EVALUATION

The proposed MInDI-3D model was compared with four state-of-the-art diffusion-based models for sparse CBCT reconstruction that include spatial information from 3 dimensions: TPDM [17], TOSM [16], DPS [40] and Blaze3DM [15]. We did not include DiffusionMBIR [18] as their implementation is focused on fewer projections. We compare the reported metrics with our own results, all of which were evaluated on comparable pseudo-CBCT datasets. As shown in Table 2, our proposed MInDI-3D model demonstrates competitive reconstruction quality compared to current state-of-the-art methods, even when using a more challenging sparse-view scenario with only 25 projections. We note that a direct comparison is challenging, as these state-of-the-art methods were evaluated on different datasets (e.g., the AAPM Low Dose CT Grand Challenge), with different projection counts, and often used 2D-averaged metrics versus our full 3D metric reporting. On the in-domain CT-RATE validation set, MInDI-3D achieves a full 3D PSNR of 36.81 dB and an SSIM of 0.95. This performance is highly competitive with methods like Blaze3DM, which reports a 38.39 dB PSNR from 36 projections using an averaged 2D metric. To provide a direct comparison, we evaluated TPDM [17] on the HyperSight test set using the same 25-projection configuration. While our model compares similarly to TPDM in terms of PSNR (29.30 dB vs. 29.38 dB), MInDI-3D outperforms TPDM significantly in terms of SSIM (0.86 vs. 0.72). This discrepancy indicates that TPDM tends to over-smooth high-frequency details, optimising pixel-wise error (PSNR) at the cost of structural fidelity, whereas MInDI-3D preserves sharp edges and fine texture, resulting in superior perceptual quality.

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).

To quantitatively evaluate anatomical consistency, we performed a task-based analysis using automated segmentation. We applied TotalSegmentator to reconstructions generated by MInDI-3D across a range of iterative refinement steps (1)–30) on the real-world HyperSight dataset. The results demonstrate that segmentation accuracy remains stable, indicating that additional refinement steps do not degrade anatomical integrity. Critical structures like lungs (DICE=0.96– 0.98), vertebrae (DICE=0.88-0.89) and heart (DICE=0.75-0.77) 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 4). 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 4). 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.

MInDI-3D achieves inference speeds competitive with the 3D U-Net baseline: a single sampling step requires 19 ms/volume versus the U-Net's 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 roughly 190 ms for model inference) and can be evaluated on a GPU with 32 GB VRAM.

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

TABLE 2. 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. Comparisons to related work are indirect, as evaluation protocols, datasets, and metrics (2D vs. 3D) differ.

Method	Projections	PSNR	SSIM	PSNR Improvement	SSIM Improvement
Related Work					
TPDM [17]	36	38.25 (2D)	0.949* (2D)	–	–
TOSM [16]	29	38.21* (2D)	0.936* (2D)	+15.56*	+0.636*
DPS [40]	30	31.29 (2D)	0.847 (2D)	+12.98	+0.617
Blaze3DM [15]	36	38.39	0.951*	–	–
Validation set (CT-RATE)					
3D U-Net (our work)	25	36.85 ± 1.23	0.94 ± 0.02	+20.24	+0.76
MInDI-3D (our work)	25	36.81 ± 1.15	0.95 ± 0.02	+20.20	+0.77
Test set (HyperSight)					
MInDI-3D (our work)	25	29.30 ± 0.95	0.86 ± 0.02	+10.00	+0.54
3D U-Net (our work)	25	28.90 ± 1.02	0.85 ± 0.02	+9.6	+0.53
TPDM [17]	25	29.38 ± 0.84	0.72 ± 0.06	+10.08	+0.40

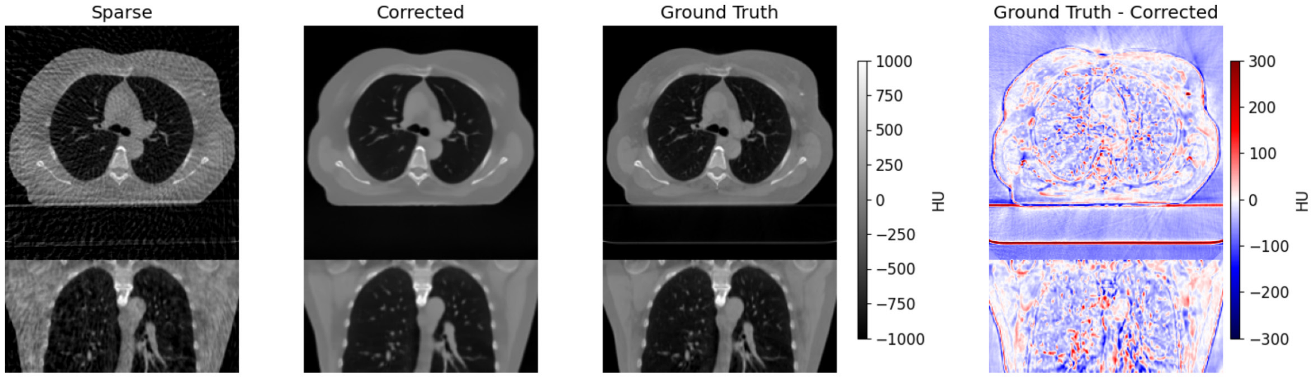


FIGURE 2. CBCT images (axial and coronal views) of a breast cancer patient (HyperSight dataset), from left to right showing the sparse volume (50 projections), corrected volume using the MInDI-3D model, ground truth volume and a difference plot (ground truth - corrected volume).

TABLE 3. MInDI-3D's 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.

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

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 number of sampling steps for fidelity, varied across images and anatomic sites.

B. 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-world HyperSight dataset (the 320-subject model was used as the 3200-subject model, was not yet available when the clinical study was conducted). Figure 2 provides a qualitative example of this correction on a breast cancer patient from the HyperSight test set. The images demonstrate a strong suppression of the streak artefacts seen in the sparse-view input, resulting in a cleaner volume. While some fine-texture smoothing is visible compared to the ground truth, key anatomical structures remain clear and well-defined. 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

TABLE 4. 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 trained on 320 subjects). 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 (Δ_{PSNR} = +2.02 dB vs. no attention blocks). Metrics averaged over test volumes versus ground truth.

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

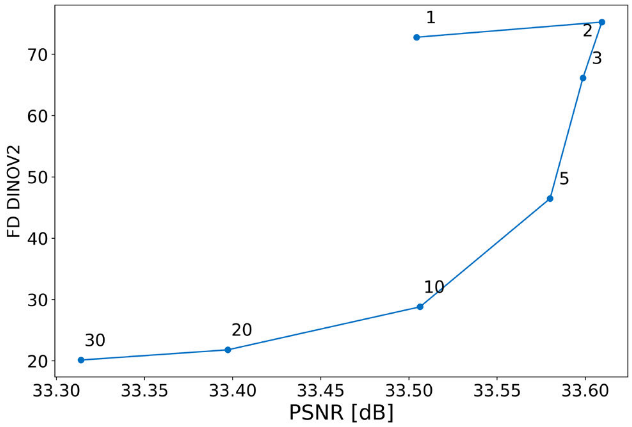


FIGURE 3. Perception-distortion trade-off in progressive sampling of MInDI-3D on the test set HyperSight with 50 projections. The line plot compares fidelity (PSNR) and perceptual quality (FD DINOv2) across sampling steps (1-10). Sampling with 2-5 steps improves distortion (higher PSNR) compared to 1 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. Note: The clinical evaluation presented in Section III-B utilized the 1-step inference model.

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 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.

IV. 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 competitive results compared with state-of-the-art models, but also offers unique advantages in terms of a tuneable image quality.

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

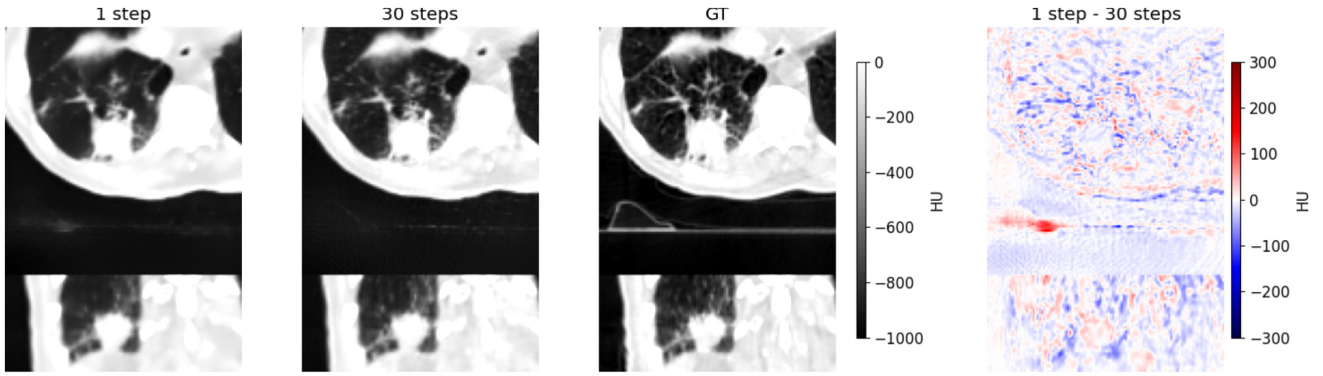


FIGURE 4. Comparing the MInDI-3D prediction of a lung tumour (lower right lung lobe) from a sparse 50 reconstruction with 1 vs. 30 steps (the ground truth and the difference of step 1 - step 30 as reference). There is an increase in sharpness and detail from step 1 to step 30.

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 lower acceptance for abdominal contouring (41.8%) likely results from the inherently lower native soft-tissue contrast in the abdomen compared to the high-contrast structures in the lung. Since MInDI-3D was trained exclusively on chest data, its learned priors may struggle to resolve subtle density gradients in the abdomen.

Our results demonstrate that MInDI-3D achieves a notable performance, particularly in its ability to reconstruct high-fidelity 3D volumes from as few as 25 projections. The +20.20 dB PSNR and +0.77 SSIM improvement over the analytical FBP reconstruction on our validation set signifies an increase in quality, outperforming the comparable models especially in terms of structural fidelity. This highlights the model's effectiveness in learning a powerful 3D prior that overcomes the severe ill-posedness of the sparse-view problem. The evaluation on the out-of-domain HyperSight test set highlights the model's generalisability. Despite the challenging domain shift, encompassing differences in scanner geometry (half-fan vs. full-fan), anatomical variance (chest-only training vs. abdomen/breast/lung testing), and acquisition physics, MInDI-3D maintains a robust performance of 29.30 dB. This indicates that the model has learned a generalised 3D prior rather than overfitting to the training distribution. Future work could further improve this cross-domain robustness by training on multi-scanner and multi-anatomy datasets or employing domain adaptation techniques. To this end, publicly available CBCT datasets that include real-world projections would be invaluable for benchmarking and enhancing performance on real-world data. The perception-distortion trade-off observed with MInDI-3D sampling steps mirrors findings in [14]. MInDI-3D users can adjust sampling steps to prioritise either quantitative fidelity or perceptual realism, tailoring the output to specific clinical needs. Based on this analysis, we propose a task-based selection strategy for clinical practice: for quantitative workflows such as dose calculation, a minimal step count (1–2 steps) is recommended to prioritise pixel fidelity (high PSNR). Conversely, for visual tasks such as manual contouring, increasing the count (e.g., to 2–5 steps) is advised to enhance perceptual definition (low FD) without significantly compromising anatomical stability. 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 to be validated for a clinical setting, building on [57]. Third, while diffusion-based models risk introducing 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. Nevertheless, future work could address this topic in greater detail through targeted radiological reviews designed to explicitly quantify subtle synthetic structures. In addition, calculating the treatment dose at different sampling steps could be a further step to validate this approach in a task-based manner.

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 voxel space, an alternative approach for future work could explore enhancement in a latent space, which through compression should allow training 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

While some authors are employed by Varian, they declare no conflicts of interest related to this work.

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," *Int. J. Radiat. Oncol.*Biol.*Phys.*, vol. 53, no. 5, pp. 1337–1349, Aug. 2002. [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*, vol. 12, no. 8, pp. 1–12, Aug. 2020. doi: 10.7759/cureus.9660.

[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," *Radiol., Imag. Cancer*, vol. 5, no. 4, Jul. 2023, Art. no. 230011, doi: 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*, vol. 12, pp. 10281–10295, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10398205

[5] 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, Switzerland: Springer, 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.

[6] 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. 21049–21067, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10056934

[7] R. Labaca-Castro, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 73–76. [Online]. Available: https://proceedings.neurips.cc/paperfiles/paper/2014/hash/f033ed80deb0234979a61f95710dbe25-Abstract.html

[8] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," Dec. 2022, *arXiv:1312.6114*.

[9] J. Ho, A. N. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2024, pp. 6840–6851. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html

[10] M. Ali et al., "Generative adversarial networks (GANs) for medical image processing: Recent advancements," in *Archives of Computational Methods in Engineering*, vol. 32, no. 2. U.K.: Springer, 2024, pp. 1185–1198.

[11] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 8780–8794. [Online]. Available: https://proceedings.neurips.cc/paperfiles/paper/2021/hash/49ad23d1ec9fa4bd8d77d02681df5cfa-Abstract.html

[12] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10674–10685. [Online]. Available: https://ieeexplore.ieee.org/document/9878449

[13] G. Müller-Franzes, L. Huck, M. Bode, S. Nebelung, C. Kuhl, D. Truhn, and T. Lemainque, "Diffusion probabilistic versus generative adversarial models to reduce contrast agent dose in breast MRI," *Eur. Radiol. Experim.*, vol. 8, no. 1, p. 53, May 2024, doi: 10.1186/s41747-024-00451-3.

[14] M. Delbracio and P. Milanfar, "Inversion by direct iteration: An alternative to denoising diffusion for image restoration," *Trans. Mach. Learn. Res.*, Mar. 2023. [Online]. Available: https://openreview.net/forum?id=VmyFF5lL3F

[15] J. He, B. Li, G. Yang, and Z. Liu, "Blaze3DM: Marry triplane representation with diffusion for 3D medical inverse problem solving," 2024, *arXiv:2405.15241*.

[16] 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," *Comput. Biol. Med.*, vol. 168, Jan. 2024, Art. no. 107819. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482523012842

[17] 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," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Paris, France, Oct. 2023, pp. 10676–10686. [Online]. Available: https://ieeexplore.ieee.org/document/10377891/

[18] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, Jun. 2023, pp. 22542–22551. [Online]. Available: https://ieeexplore.ieee.org/document/10204965/

[19] Y. Blau and T. Michaeli, "The perception-distortion tradeoff," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6228–6237.

[20] R. Schulze, U. Heil, D. Groß, D. Bruellmann, E. Dranischnikow, U. Schwanecke, and E. Schoemer, "Artefacts in CBCT: A review," *Dentomaxillofacial Radiol.*, vol. 40, no. 5, pp. 265–273, Jul. 2011, doi: 10.1259/dmfr/30642039.

[21] F. E. Boas and D. Fleischmann, "CT artifacts: Causes and reduction techniques," *Imag. Med.*, vol. 4, no. 2, pp. 229–240, Apr. 2012. [Online]. Available: https://www.openaccessjournals.com/abstract/ct-artifacts-causes-and-reduction-techniques-9353.html

[22] M. Amirian, J. A. Montoya-Zegarra, I. Herzig, P. Eggenberger Hotz, L. Lichtensteiger, M. Morf, A. Züst, P. Paysan, I. Peterlik, S. Scheib, R. M. Füchslin, T. Stadelmann, and F.-P. Schilling, "Mitigation of motion-induced artifacts in cone beam computed tomography using deep convolutional neural networks," *Med. Phys.*, vol. 50, no. 10, pp. 6228–6242, Oct. 2023, doi: 10.1002/mp.16405.

[23] 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, Switzerland: Springer, 2023, pp. 86–100, doi: 10.1007/978-3-031-26351-4_6.

[24] D. Hu, Y. Zhang, J. Liu, Y. Zhang, J. L. Coatrieux, and Y. Chen, "PRIOR: Prior-regularized iterative optimization reconstruction for 4D CBCT," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 11, pp. 5551–5562, Nov. 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9866113

[25] 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," *Quant. Imag. Med. Surgery*, vol. 11, no. 12, pp. 4767–4780, Dec. 2021. [Online]. Available: https://qims.amegroups.org/article/view/75305

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

[27] S. Kim et al., "A 3D conditional diffusion model for image quality transfer—An application to low-field MRI," in *Deep Generative Models for Health Workshop NeurIPS*, Oct. 2023. [Online]. Available: https://openreview.net/forum?id=TynSiNAVc8

[28] Y. Liu, G. Dwivedi, F. Boussaid, F. Sanfilippo, M. Yamada, and M. Bennamoun, "Inflating 2D convolution weights for efficient generation of 3D medical images," *Comput. Methods Programs Biomed.*, vol. 240, Oct. 2023, Art. no. 107685. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169260723003504

[29] 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, Switzerland: Springer, 2020, pp. 657–666, doi: 10.1007/978-3-030-59728-3_64.

[30] J. Kapoor, J. H. Macke, and C. F. Baumgartner, ''Multiscale metamorphic VAE for 3D brain MRI synthesis,'' 2023, *arXiv:2301.03588*.

[31] P. Friedrich, J. Wolleb, F. Bieder, A. Durrer, and P. C. Cattin, ''WDM: 3D wavelet diffusion models for high-resolution medical image synthesis,'' in *Deep Generative Models*, A. Mukhopadhyay, I. Oksuz, S. Engelhardt, D. Mehrof, and Y. Yuan, Eds., Cham, Switzerland: Springer, 2025, pp. 11–21, doi: 10.1007/978-3-031-72744-3_2.

[32] 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,'' *Med. Phys.*, vol. 51, no. 4, pp. 2538–2548, Apr. 2024, doi: 10.1002/mp.16847.

[33] Z. Dorjsembe, H.-K. Pao, S. Odonchimed, and F. Xiao, ''Conditional diffusion models for semantic 3D brain MRI synthesis,'' *IEEE J. Biomed. Health Informat.*, vol. 28, no. 7, pp. 4084–4093, Jul. 2024.

[34] 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,'' *Med. Image Anal.*, vol. 91, Jan. 2024, Art. no. 102983. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841523002438

[35] S. Poonkodi and M. Kanchana, ''3D-MedTranCSGAN: 3D medical image transformation using CSGAN,'' *Comput. Biol. Med.*, vol. 153, Feb. 2023, Art. no. 106541. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482523000069

[36] 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, Switzerland: Springer, 2025, pp. 255–278, doi: 10.1007/978-3-031-80965-1_13.

[37] 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 *Proc. 45th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2023, pp. 1–4. [Online]. Available: https://ieeexplore.ieee.org/document/10341112

[38] 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, Switzerland: Springer, 2022, pp. 516–526, doi: 10.1007/978-3-031-16446-0_49.

[39] C. McCollough, ''TU-FG-207A-04: Overview of the low dose CT grand challenge,'' *Med. Phys.*, vol. 43, no. 6, pp. 3759–3760, Jun. 2016, doi: 10.1118/1.4957556.

[40] 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,'' *J. Med. Imag.*, vol. 11, no. 4, Aug. 2024, Art. no. 043504.

[41] H. R. Roth et al., ''A new 2.5D representation for lymph node detection using random sets of deep convolutional neural network observations,'' in *Medical Image Computing and Computer-Assisted Intervention: MICCAI*. Cham, Switzerland: Springer, 2014, pp. 520–527, doi: 10.1007/978-3-319-10404-1_65.

[42] 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. Özdemir, and B. H. Menze, ''A foundation model utilizing chest CT volumes and radiology reports for supervised-level zero-shot detection of abnormalities,'' 2024, *arXiv.2403.17834*.

[43] L. A. Feldkamp, L. C. Davis, and J. W. Kress, ''Practical cone-beam algorithm,'' *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 1, no. 6, pp. 612–619, Jun. 1984. [Online]. Available: https://opg.optica.org/josaa/abstract.cfm?uri=josaa-1-6-612

[44] A. Andersen, ''Simultaneous algebraic reconstruction technique (SART): A superior implementation of the ART algorithm,'' *Ultrason. Imag.*, vol. 6, no. 1, pp. 81–94, Jan. 1984. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0161734684900087

[45] G. N. Ramachandran and A. V. Lakshminarayanan, ''Three-dimensional reconstruction from radiographs and electron micrographs: Application of convolutions instead of Fourier transforms,'' *Proc. Nat. Acad. Sci. USA*, vol. 68, no. 9, pp. 2236–2240, Sep. 1971, doi: 10.1073/pnas.68.9.2236.

[46] Ö. Çiçek, 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, Switzerland: Springer, 2016, pp. 424–432, doi: 10.1007/978-3-319-46723-8_49.

[47] S. Ioffe and C. Szegedy, ''Batch normalization: Accelerating deep network training by reducing internal covariate shift,'' in *Proc. 32nd Int. Conf. Mach. Learn.*, Jun. 2015, pp. 448–456. [Online]. Available: https://proceedings.mlr.press/v37/ioffe15.html

[48] D. Hendrycks and K. Gimpel, ''Gaussian error linear units (GELUs),'' 2023, *arXiv:1606.08415*.

[49] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, ''Attention is all you need,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2025, pp. 5998–6008. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[50] P. Esser, R. Rombach, and B. Ommer, ''Taming transformers for high-resolution image synthesis,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12868–12878. [Online]. Available: https://ieeexplore.ieee.org/document/9578911

[51] D. P. Kingma and J. Ba, ''Adam: A method for stochastic optimization,'' 2014, *arXiv:1412.6980*.

[52] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, ''Image quality assessment: From error visibility to structural similarity,'' *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004. [Online]. Available: https://ieeexplore.ieee.org/document/1284395

[53] N. Otsu, ''A threshold selection method from gray-level histograms,'' *IEEE Trans. Syst. Man, Cybern.*, vol. SMC-9, no. 1, pp. 62–66, Jan. 1979. [Online]. Available: https://ieeexplore.ieee.org/document/4310076

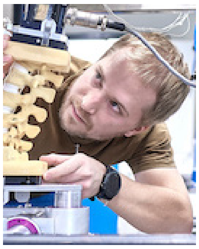
[54] G. Stein, J. C. Cresswell, R. Hosseinzadeh, Y. Sui, B. L. Ross, V. Villecroze, Z. Liu, A. L. Caterini, J. Taylor, and G. Loaiza-Ganem, ''Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models,'' in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2023, pp. 3732–3784. [Online]. Available: https://proceedings.neurips.cc/paperfiles/paper/2023/hash/0bc795afae289ed465a65a3b4b1f4eb7-Abstract-Conference.html

[55] M. Oquab et al., ''DINOv2: Learning robust visual features without supervision,'' *Trans. Mach. Learn. Res. J.*, 2023. [Online]. Available: https://openreview.net/forum?id=VmyFF5lL3F

[56] 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,'' *Radiol., Artif. Intell.*, vol. 5, no. 5, Sep. 2023, Art. no. 230024, doi: 10.1148/ryai.230024.

[57] M. Woodland, A. Castelo, M. Al Taie, J. A. M. Silva, M. Eltaher, F. Mohn, A. Shieh, S. Kundu, J. P. Yung, A. B. Patel, and K. K. Brock, ''Feature extraction for generative medical imaging evaluation: New evidence against an evolving trend,'' in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2024*, M. G. Linguraru, Ed., Cham, Switzerland: Springer, 2024, pp. 87–97, doi: 10.1007/978-3-031-72390-2_9.

DANIEL BARCO received the M.Sc. degree in applied information and data science from Lucerne University of Applied Sciences and Arts, in 2020. He is currently pursuing the Ph.D. degree in AI with the University of Zurich (UZH). He is a Researcher with the Centre for Artificial Intelligence, Zurich University of Applied Sciences (ZHAW), Winterthur, Switzerland. His research interests include pioneering novel neural architectures for computer vision, while also contributing to the development of robust and trustworthy AI solutions.



MARC STADELMANN received the M.Sc. degree in biomedical engineering and the Ph.D. degree in computational biomechanics from the University of Bern, Switzerland, in 2013 and 2018, respectively. He is currently a Researcher with the ZHAW Centre for Artificial Intelligence (CAI) and part of the Intelligent Vision Systems Group. His research interests include computer vision and specifically medical imaging.

MARTIN OSWALD received the B.Sc. degree in computer science from the ZHAW. He is currently pursuing the master's degree. He was the recipient of the Audience Award at the National Siemens Excellence Award, in 2024 for his bachelor's thesis, which developed a deep learning model for the detection of thyroid cancer. His research interests include implementing and optimizing deep learning networks to solve clinical problems.

IVO HERZIG is an Engineer and a Computer Scientist with professional background in software development, computational geometry, computer graphics and robotics. He is currently a Researcher with the Institute of Applied Mathematics and Physics (IAMP), ZHAW, Switzerland, where he is focusing on deep learning for medical image analysis in the area of image-guided radiation therapy (IGRT).

LUKAS LICHTENSTEIGER received the M.Sc. degree in theoretical physics and the Ph.D. degree in AI and robotics from the University of Zurich, in 1995 and 2004, respectively. He is currently a Lecturer with Zurich University of Applied Sciences (ZHAW), School of Engineering, where he is heading the Medical Complex Systems Group. His research interests include applying advanced computational methods and physics-based modeling to medical image reconstruction and analysis.

PASCAL PAYSAN received the Ph.D. degree in computer science from the University of Basel, Switzerland, in 2010. He is currently a Senior Staff Research Scientist with the Varian Medical Systems Imaging Laboratory, Baden, Switzerland. With over a decade of experience at Varian, he has served as the Tech Lead for the Cone-Beam CT Reconstruction Framework. His research interests include CBCT artifact reduction, iterative and statistical reconstruction methods, and motion compensation for image-guided radiation therapy.

IGOR PETERLIK received the Ph.D. degree in informatics from Masaryk University, Czech Republic, in 2009. He is currently a Senior Research Scientist with Varian, a Siemens Healthineers Company. Previously, he held a tenured research position with Inria, French National Institute for Research in Digital Science and Technology. He completed Postdoctoral Fellowships with INRIA Lille and the University of British Columbia. His research interests include patient-specific biomechanical modeling, real-time simulation of deformable objects, and medical image reconstruction.

MICHAL WALCZAK received the master's and Ph.D. degrees in physics from The University of Göttingen, Germany, in 2010 and 2014, respectively. He was a Postdoctoral Researcher with the Max Planck Institute for Biophysical Chemistry, developing pattern recognition algorithms. Subsequently, he was a Research Fellow with the Fraunhofer Institute for Industrial Mathematics ITWM, where his work included software development for interactive radiotherapy planning and machine learning. Since July 2020, he has been a Research Scientist with Varian Medical Systems Imaging Laboratory, Baden, Switzerland, contributing to the development of deep learning models for medical imaging.

BJOERN MENZE received the degree in physics from the Universities of Heidelberg, Germany, and Uppsala, Sweden, and the Ph.D. degree in computer science from Heidelberg University, in 2007. He held Postdoctoral positions with INRIA, Sophia-Antipolis, ETH Zurich, Harvard University, and MIT. He was a Professor with the Technical University of Munich (TUM). He is currently a Full Professor of the biomedical image analysis and machine learning with the University of Zurich (UZH). His research interests include medical image computing and explores topics at the interface of medical computer vision, image-based modeling, and computational physiology, with applications in clinical neuroimaging and the modeling of tumor growth.

FRANK-PETER SCHILLING received the Ph.D. degree in physics from the University of Heidelberg, Germany, in 2001. He subsequently spent many years in fundamental research at physics laboratories including CERN, Geneva, Switzerland, where he was involved in the discovery of the Higgs particle, in 2012. Besides managing international scientific projects and teams, and being a top-cited author of particle physics research journal publications (h-index of 150), he developed a strong profile in computer science, big data, statistical modelling, and machine learning. He joined Zurich University of Applied Sciences ZHAW, Winterthur, Switzerland, in 2018, and he is a Senior Lecturer, the Group Leader, and the Deputy Head of ZHAW's Centre for AI (CAI). His research interests include AI and deep learning, with a focus on computer vision (in particular for medical imaging), as well as on machine learning operations (MLOps). In addition, he is interested in trustworthy and certifiable AI, as well as in applications of deep learning in the physical sciences.