IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 15 December 2025, accepted 5 January 2026, date of publication 12 January 2026, date of current version 27 January 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3653234

RESEARCH ARTICLE

Cone-Beam CT to Synthetic CT Translation Using Conditional 3D Latent Diffusion-Based Model

MOHAMMED AL-SHALABI[1], MOHAMMED A. MAHDI[1], REDA ELBAROUGY[2], EHAB TAWFEEK ALNFRAWY[3], MUHAMMAD USMAN HADI[4], (Member, IEEE), AND RAO FAIZAN ALI[5], (Member, IEEE)

[1]Information and Computer Science Department, College of Computer Science and Engineering, University of Hail, Ha'il 55476, Saudi Arabia
[2]Artificial Inelegance and Data Science Department, College of Computer Science and Engineering, University of Hail, Ha'il 55476, Saudi Arabia
[3]Information Security Department, College of Computer Science and Engineering, University of Hail, Ha'il 55476, Saudi Arabia
[4]School of Engineering, Ulster University, BT15 1AP Belfast, U.K.
[5]School of Computing, University of Kent, CT2 7NZ Canterbury, U.K.

Corresponding author: Rao Faizan Ali (R.F.Ali@kent.ac.uk)

This work was supported by the Scientific Research Deanship at University of Hail, Saudi Arabia, under Project RG-24 182.

ABSTRACT Accurate cone-beam CT (CBCT)-to-synthetic CT (sCT) translation is essential for image-guided adaptive radiotherapy (IGART), where Hounsfield unit (HU) fidelity and structural accuracy directly affect dose calculation. We propose a conditional 3D Latent Diffusion Model (3DLDFM) for volumetric CBCT-to-sCT synthesis. The framework comprises two stages: 1) a 3D variational autoencoder with KL regularization that compresses CBCT volumes into a three-channel latent representation, trained with a composite loss combining L1 reconstruction, perceptual, KL, and adversarial terms; and 2) a conditional 3D U-Net diffusion model that performs iterative denoising in latent space using a DDPM-style noise schedule, conditioned on the input CBCT. We evaluated 3DLDFM on the multi-center SynthRAD2023 dataset comprising 955 paired CBCT/CT volumes spanning head-and-neck, thorax, and abdominal sites. Performance is benchmarked against SwinUNETR, nnUNet, CycleGAN, and Pix2Pix using Mean Absolute Error (MAE), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index Measure (SSIM) within body masks. Across all regions, 3DLDFM achieves the lowest overall MAE (51.40 HU) and the highest overall SSIM (0.9124), while maintaining competitive PSNR (30.60 dB), surpassing all baselines in HU accuracy and structural fidelity. These results demonstrate that the proposed latent diffusion framework provides a robust and generalizable solution for CBCT-to-CT synthesis and strengthens the feasibility of simulation-free adaptive radiotherapy workflows.

INDEX TERMS Cone-beam CT, synthetic CT, latent diffusion model, image-guided adaptive radiotherapy, image-to-image translation, deep learning.

I. INTRODUCTION

Image-Guided Adaptive Radiotherapy (IGART) represents the current frontier in precision oncology, offering the ability to dynamically modify a patient's treatment plan in response to daily anatomical changes, thereby maximizing therapeutic dose to the target volume while minimizing toxicity to organs-at-risk [1], [2], [3], [4]. A foundational challenge in implementing IGART is the accurate and rapid recalculation of the radiation dose on a daily basis [5], [6]. While Cone Beam Computed Tomography (CBCT) is the standard on-board imaging modality, providing essential high-resolution geometric information for daily patient setup and tracking, its clinical utility for quantitative dose calculation is severely limited [7], [8], [9]. CBCT scans inherently suffer from significant artifacts due to scatter, image noise, and beam hardening, resulting in substantial Hounsfield Unit (HU) inaccuracies that can lead to dosimetric errors of up to 5-10%—a margin unacceptable for high-precision adaptive planning [10]. Consequently, the clinical workflow demands

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

a robust method to synthesize a high-quality, planning synthetic CT (sCT) image from the daily CBCT scan, effectively correcting the HU values to enable accurate dose accumulation and real-time adaptation [11], [12].

The critical need for high-fidelity CBCT-to-sCT generation has spurred extensive research, particularly leveraging advancements in Deep Learning (DL) [5], [13], [14]. Early approaches based on deformable image registration (DIR) struggle with complex tissue interfaces and large anatomical shifts [15], [16]. The field quickly transitioned to Convolutional Neural Networks (CNNs), prominently featuring U-Net and Generative Adversarial Networks (GANs) [17], [18], [19], [20], [21], [22], which model the image-to-image translation task. While these methods have achieved clinical feasibility by significantly mitigating artifacts and improving HU accuracy, they face two key, persistent limitations that hinder their widespread adoption in high-stakes IGART. Firstly, models trained with mean-squared error (L1/L2) loss often yield over-smoothed sCT outputs, sacrificing the fine textural details and sharp edges critical for distinguishing between soft tissue structures and precisely localizing heterogeneous bony interfaces [22], [23]. Secondly, their deterministic nature struggles to capture the inherent conditional image distribution, leading to limited generalization when encountering novel artifacts or anatomical variances outside the training distribution, which are common in clinical practice [24], [25].

Despite these advances, current CBCT-to-sCT methods still face a fundamental tension between HU fidelity, structural sharpness, and robustness across scanners, anatomies, and institutions. In the context of IGART, where inaccurate HUs can propagate directly into erroneous dose calculations and suboptimal plan adaptations, there is a pressing need for generative models that (i) explicitly model the full 3D volumetric structure, (ii) remain resilient to CBCT-specific noise and artifacts, and (iii) are computationally tractable for routine, daily clinical use. These requirements motivate the development of a principled, probabilistic framework that can better capture the underlying conditional distribution of sCT given CBCT, rather than relying solely on deterministic pixel-space mappings.

To address these shortcomings and elevate the quality and reliability of sCT generation, this paper introduces an approach based on the Three-Dimensional Latent Diffusion Model (3D LDM). Diffusion Models, characterized by an iterative denoising process, have demonstrated unparalleled performance in generating high-quality, photorealistic images by accurately modeling complex data distributions [24]. By implementing the diffusion process in a compressed latent space, the 3D LDM architecture efficiently handles the substantial computational requirements of volumetric medical data (CBCT and sCT) while effectively conditioning the generation on the CBCT volume. This strategic use of 3D architecture ensures that cross-slice anatomical consistency and full volumetric context are preserved during the translation, a crucial factor often compromised by slice-by-slice 2D methods. The main contributions of this study are summarized as follows:

- We propose a conditional Three-Dimensional Latent Diffusion Model (3DLDFM) for CBCT-to-sCT synthesis that operates in a compressed latent space, enabling computationally efficient, fully volumetric generative modeling while preserving cross-slice anatomical consistency.
- We designed a composite training objective for the autoencoder that combines L1 reconstruction, perceptual, KL divergence, and adversarial losses, explicitly targeting HU fidelity, structural sharpness, and perceptual realism in the reconstructed sCT volumes.
- We conduct a comprehensive evaluation on the SynthRAD2023 dataset comprising 955 paired CBCT/CT volumes across head-and-neck, thorax, abdominal sites, and benchmark 3DLDFM against strong baselines (SwinUNETR, nnUNet, CycleGAN, Pix2Pix) using MAE, PSNR, and SSIM.
- We demonstrate that 3DLDFM consistently improves HU accuracy and structural similarity over all baseline models and anatomical regions while maintaining competitive PSNR, thereby providing a robust and generalizable solution for simulation-free IGART workflows.

The structure of this paper is as follows: Section II reviews the background of image-to-image translation in radiotherapy and the emergence of diffusion models. Section III elaborates on the technical details of the proposed 3D Latent Diffusion Model architecture. Section IV describes the experimental methodology, dataset, and training protocol. Section V presents and discusses the superior performance achieved in both image quality and dosimetry accuracy. Finally, Section VI concludes the paper and outlines future research directions.

II. RELATED WORK

Generating sCT from CBCT has become a central enabler for image-guided adaptive radiotherapy, where accurate HU fidelity and robust artifact suppression are prerequisites for daily dose recalculation and replanning [3], [26]. Recent surveys emphasize that sCT pipelines have matured substantially and are increasingly integrated into adaptive workflows. However, unresolved challenges remain, including scanner- and anatomy-specific domain shifts, HU accuracy for reliable dose computation, and generalization under limited paired data availability [7], [24].

Early deep-learning approaches were dominated by supervised convolutional encoder–decoder architectures, most prominently 2D and 3D U-Nets, which leveraged paired CBCT–planning CT (pCT) volumes to reduce HU errors and improve image similarity metrics such as PSNR and SSIM [22], [27], [28]. These models consistently outperformed raw CBCT in both image quality and dosimetric accuracy, with reports documenting improved organ-at-risk sparing and tumor dose coverage on recalculated treatment plans [5], [27]. Despite these successes, their sensitivity to

acquisition-specific artifacts, such as scatter, motion, and truncation, as well as the need for large, curated datasets, limited their robustness in clinical deployment.

To address the limitations of paired training, unpaired adversarial models, particularly CycleGAN and its derivatives, were rapidly adapted for CBCT-to-sCT translation. These approaches achieved encouraging gains in structural fidelity and artifact suppression without requiring paired training, including applications in head-and-neck and pediatric cohorts [11]. More recent work has introduced Vision Transformer components into adversarial pipelines, overcoming CycleGAN's local-context bias and enabling better HU stability and structural consistency [17], [29]. Unified multi-site frameworks have also been proposed to improve generalizability across anatomies and scanners, a prerequisite for routine IGART deployment [30]. Yet, despite advances in perceptual realism, these architectures remain constrained by adversarial training instabilities and limited HU calibration, both of which restrict their reliability for downstream dose evaluation.

More recently, denoising diffusion probabilistic models (DDPMs) have emerged as a state-of-the-art alternative for medical image translation and CBCT enhancement. Diffusion models iteratively refine noisy inputs, offering stronger mode coverage, superior edge preservation, and reduced risk of hallucinated structures compared to adversarial methods. Applied to CBCT-to-CT synthesis, conditional diffusion models have demonstrated marked improvements in HU fidelity and structural preservation relative to CNN- and GAN-based baselines [22]. Landmark studies in 2023–2024 validated diffusion methods on thoracic and multi-site datasets, confirming their ability to generalize across anatomies while retaining voxel-wise accuracy [31], [32]. Advanced diffusion variants have further introduced adaptive high-frequency optimization and hybrid U-Net–ViT backbones, enabling recovery of fine trabecular details essential for dose recalculation and contouring [33], [34]. Other extensions address sparse-view or limited-angle CBCT using frequency-guided priors, illustrating the flexibility of diffusion frameworks for diverse IGART acquisition regimes [35].

Progress in this field has been accelerated by the release of large-scale benchmarks such as SynthRAD2023, which provided harmonized evaluation protocols across MRI–CT and CBCT–CT tracks, spanning over a thousand subjects [7], [8]. Standard evaluation metrics, HU-MAE/RMSE, PSNR, SSIM, and clinical endpoints such as γ-pass rates and dose–volume indices have enabled transparent comparisons across methods. Recent reviews show consistent improvements for CBCT-derived sCT, with diffusion-based methods increasingly leading quantitative benchmarks [5], [27].

From a translational perspective, clinical studies underscore the need to balance image quality with workflow efficiency and imaging dose. Simulation-free or CBCT-driven replanning pipelines now demonstrate reduced margins and improved organ-at-risk protection, provided that image quality and HU fidelity are sufficient [1], [2], [3]. Diffusion-based synthesis is particularly well aligned with these requirements, as its stable training dynamics, fine-detail recovery, and linear HU fidelity directly support daily replanning.

III. METHODOLOGY

The proposed framework employs a two-stage three-dimensional latent diffusion model to generate sCT volumes from input CBCT scans, as depicted in **Figure 1**. By operating in a compressed latent space, the model substantially reduces computational overhead while preserving fine structural and intensity details critical for IGART. The pipeline encompasses data preprocessing, latent diffusion-based translation, and rigorous evaluation against state-of-the-art baselines, each described in detail in the subsequent sections.

A. DATASET

This study makes use of a multi-center CBCT–CT dataset collected from three academic hospitals: UMC Utrecht, UMC Groningen, and Radboud University Medical Center (Nijmegen, The Netherlands) as reported by [7]. The dataset comprising a total of 955 paired volumes acquired from patients undergoing radiotherapy. It includes three anatomical regions frequently encountered in adaptive radiotherapy workflows: head and neck (325 cases), thorax (321 cases), and abdomen (309 cases). Each patient record contains a CBCT acquired during treatment and a corresponding planning CT that has been rigidly aligned to the CBCT frame, enabling voxel-level supervision for synthetic CT generation. The dataset was divided into training and test cohorts on a patient basis to prevent data leakage across splits. Specifically, the training set consists of 765 CBCT–CT pairs (260 head and neck, 257 thorax, 248 abdomen), while the test set includes 190 pairs (65 head and neck, 64 thorax, 61 abdomen). This yields an approximate 80%–20% split between training and test data. No patient overlap exists between the splits, ensuring unbiased evaluation.

TABLE 1. Summary of the dataset by anatomical site and split.

Task	Head and Neck	Thorax	Abdominal	All
Train	260	257	248	765
Test	65	64	61	190
All	325	321	309	955

B. PREPROCESSING

To ensure reproducibility and optimize downstream learning, all CBCT and CT volumes underwent a standardized preprocessing pipeline, extending beyond the baseline steps provided in the SynthRAD2023 challenge (file conversion, resampling, image registration, and anonymization). Our pipeline was specifically designed to reduce inter-patient variability, stabilize intensity distributions, and minimize computational overhead during training. First, anatomical orientation was standardized to the radiological RAS convention, ensuring spatial consistency across all datasets.

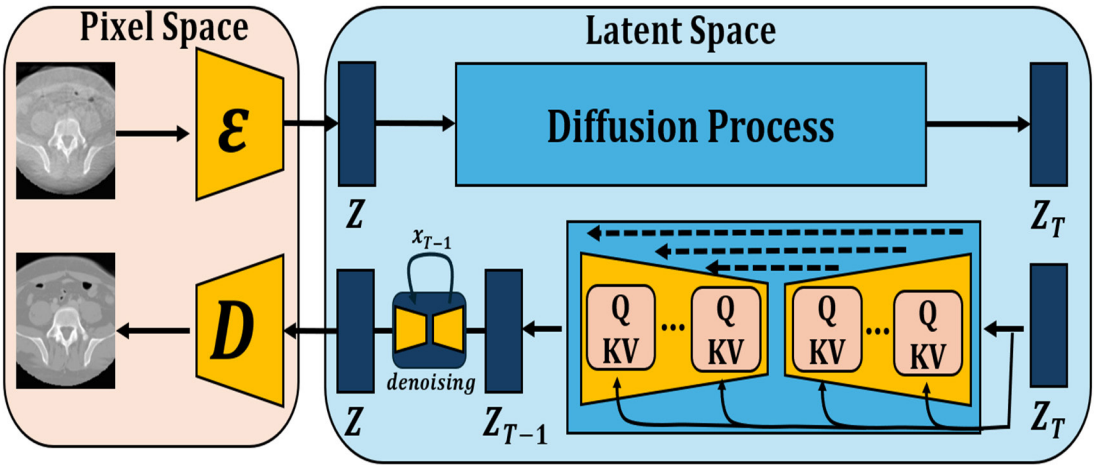

FIGURE 1. Schematic representation of the proposed 3DLDFM for CBCT-to-CT synthesis.

Intensity normalization was then performed by clipping voxel values to each volume, thereby suppressing extreme outliers from scatter or truncation artifacts. The resulting intensities were subsequently rescaled to a fixed [0, 1] range to harmonize voxel intensities across patients. To achieve spatial uniformity, all volumes were resampled to an isotropic resolution of (2.4 × 2.4 × 2.2) mm using bilinear interpolation. This step balanced computational efficiency with sufficient preservation of anatomical detail for accurate synthesis. Finally, a center crop of (96 × 96 × 64) voxels was applied, focusing the model's attention on the relevant anatomical field of view while discarding redundant background regions. This tailored preprocessing ensured that the latent diffusion model operated on a consistent and computationally tractable representation, while retaining the structural and intensity fidelity necessary for robust CBCT-to-sCT translation.

C. PROPOSED MODEL

Our proposed framework employs a 3D Latent Diffusion Model (3D-LDM) to synthesize high-quality synthetic CT (sCT) volumes from daily CBCT scans. The pipeline is designed to address two key challenges in adaptive radiotherapy: (i) the need for computationally efficient volumetric modeling, and (ii) the preservation of HU fidelity and anatomical detail in the presence of CBCT-specific artifacts. To this end, the architecture integrates two synergistic components: (1) a variational autoencoder with KL regularization (AutoencoderKL) for latent space compression, and (2) a DiffusionModelUNet that operates exclusively in the compressed representation to model the conditional distribution of CT from CBCT input. The overall workflow is illustrated in **Figure 1** and a more detailed architecture of the proposed model shown in **Figure 2**.

In our implementation, each rigidly aligned CBCT/CT pair is resampled to a uniform voxel spacing and center-cropped to a 3D volume of size 96 × 96×64. The AutoencoderKL then compresses this volume into a latent tensor with three channels, which is used as the working space for the latent diffusion process. All 3D convolutions in both the AutoencoderKL and the DiffusionModelUNet use 3 × 3×3 kernels; downsampling in the encoder path is implemented with stride 2, while all other convolutions use stride 1. We employ the SiLU (Swish) activation function in all intermediate layers and omit the activation in the final output layer of the decoder to allow direct regression of HU values.

1) LATENT SPACE COMPRESSION VIA AUTOENCODERKL

The AutoencoderKL module converts the high-dimensional volumetric data into a low-dimensional latent representation. This reduces the computational burden of 3D diffusion modeling while retaining the anatomical fidelity required for dose calculations. The autoencoder consists of a 3D encoder-decoder architecture with input/output channels of size one, and internal feature maps configured at [32, 64, 64]. The encoder (E) maps the CBCT into a latent tensor z, while the decoder (D) reconstructs the sCT from this latent space. The key hyperparameters of the proposed 3D LDM listed in Table 2.

The latent distribution is parameterized by a mean μ and variance σ^2, enabling stochastic sampling of z through the reparameterization trick. This VAE-style formulation prevents degenerate representations and improves generalization. By compressing volumetric scans into a compact latent space, the model achieves a favorable trade-off between fidelity and efficiency, enabling the subsequent diffusion stage to be trained on full 3D volumes. In particular, the encoder applies three successive stride-2 3D convolutions to reduce the spatial resolution, while the decoder mirrors this structure with transposed 3D convolutions to recover the original volume size, ensuring that cross-slice anatomical context is propagated through the latent representation.

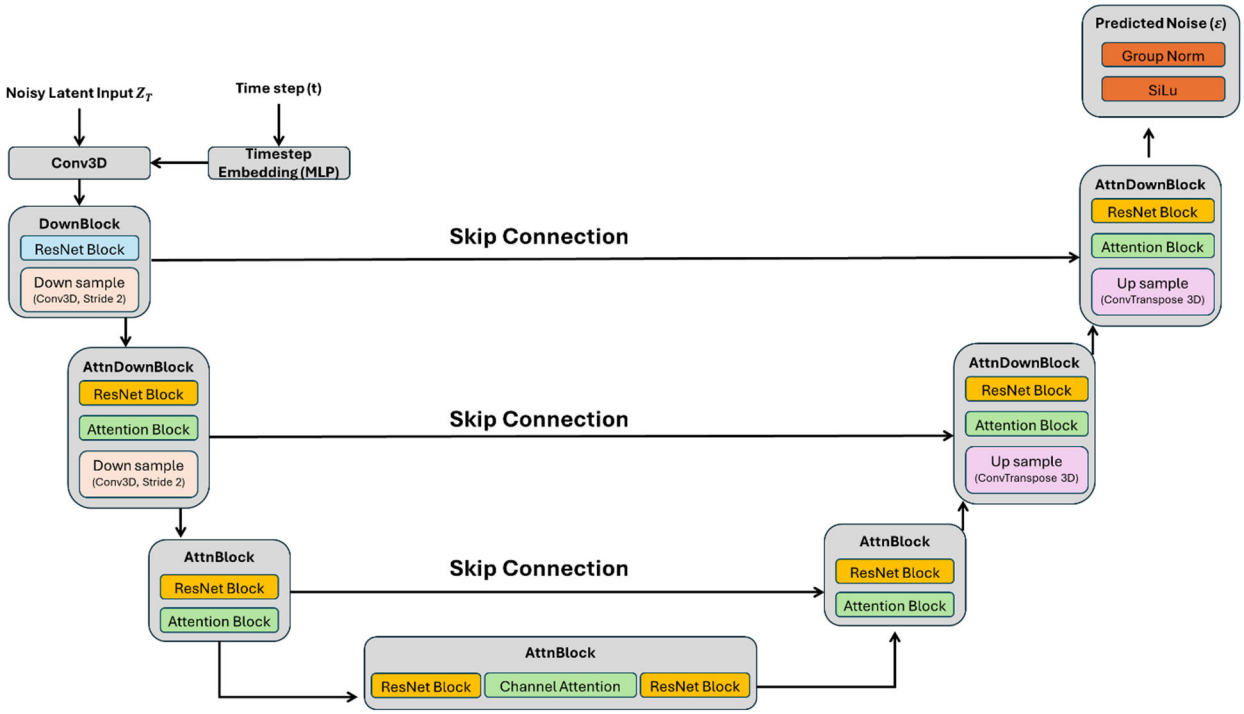

FIGURE 2. A details architecture of the of the proposed 3DLDFM.

2) DIFFUSION MODEL IN LATENT SPACE

The generative core of our framework is a DiffusionModelUNet, which reverses the gradual noising process characteristic of diffusion models. Unlike pixel-space diffusion, the model operates entirely in the compressed latent space, dramatically improving computational tractability for volumetric CBCT data.

The UNet accepts three input and three output channels corresponding to the latent dimensionality, with convolutional filters of size [32, 64, 64]. Self-attention is incorporated at deeper layers to enhance global context modeling [36]. During training, a forward diffusion process iteratively corrupts the latent vector with Gaussian noise over 1,000 steps, governed by a scaled linear beta schedule ($\beta_{\text{start}} = 0.0015$, $\beta_{\text{end}} = 0.0195$). During inference, the process is reversed, allowing the model to denoise from pure Gaussian noise to generate realistic sCT volumes conditioned on CBCT input.

3) LOSS FUNCTIONS AND TRAINING STRATEGY

The training procedure for the proposed framework consists of two complementary stages, each designed with objectives suited to its role in the pipeline. In the first stage, the autoencoder is trained to compress CBCT volumes into a compact yet information-preserving latent space. To achieve this, we employ a composite objective that balances pixel-level accuracy, perceptual fidelity, latent regularization, and adversarial realism. The reconstruction loss enforces voxel-wise fidelity between the original CBCT and its reconstruction using an L1 distance as in (1):

$$L_{\text{recon}} = \|x - \mathcal{D}(E(x))\|_1, \tag{1}$$

where x is the original image, and $\mathcal{D}(\mathcal{E}(x))$ is its reconstruction through the encoder-decoder pathway, and $\|\cdot\|_1$ is again the voxel-wise L1 norm applied to feature maps. To further encourage preservation of high-level structural features, a perceptual loss is introduced as in (2):

$$L_p = \|\psi(x) - \psi(\mathcal{D}(\mathcal{E}(x)))\|_1, \tag{2}$$

where ψ denotes a pretrained SqueezeNet feature extractor. The L1 norm was deliberately chosen over the L2 norm as it offers greater robustness against severe CBCT artifacts (outliers) and is known to produce sharper images by avoiding the over-smoothing tendency inherent to L2-based optimization, thereby preserving the fine structural details critical for accurate tissue segmentation. To regularize the latent representation, a KL divergence term aligns the encoded latent distribution with a unit Gaussian as in (3):

$$L_{kl} = \frac{1}{2}\sum_i \left(\mu_i^2 + \sigma_i^2 - \log\left(\sigma_i^2\right) - 1\right), \tag{3}$$

where μ and σ^2 represent the mean and variance of the latent distribution, the index i runs over all latent channels and spatial locations. Finally, an adversarial loss based on the least-squares objective encourages perceptual realism as in (4):

$$L_{adv} = \mathbb{E}\left[(D(\hat{x}) - 1)^2\right] \tag{4}$$

TABLE 2. Key hyperparameters of the proposed 3D LDM.

Component	Hyperparameter	Value / Setting
Input Data	Spatial size (cropped)	96*96*64 voxels
	Batch size	2
AutoencoderKL	Feature channels	[32, 64, 64]
	Latent channels	3
	Reconstruction loss	L1 Loss
	Perceptual loss weight (W_p)	0.001
	KL loss weight (W_{KL})	10^{-6}
	Adversarial loss weight (W_{adv})	0.01
	Optimizer	Adam
	Learning rate	10^{-4}
	Training epochs	100 (5 warmup epochs)
Discriminator	Architecture	PatchDiscriminator (3 layers)
	Base channels	32
	Optimizer	Adam
	Learning rate	10^{-4}
Diffusion Model	Architecture	DiffusionModelUNet
	Feature channels	[32, 64, 64]
	Attention levels	[False, True, True]
	Num. ResNet blocks	1
	Activation function	SiLU (Swish)
	Optimizer	Adam
	Learning rate	10^{-4}
	Training epochs	150
Noise Scheduler	Type	DDPMScheduler
	Timesteps (T)	1000
	Schedule type	Scaled linear beta
	Beta start / end	0.0015/0.0195

where $D(\hat{x})$ is the discriminator's prediction for a reconstructed image. Here, $\mathbb{E}[\cdot]$ denotes expectation over the mini-batch of reconstructed samples $\hat{x} = D(E(x))$. The total autoencoder loss is then expressed as in (5):

$$L_{\text{autoencoder}} = L_{\text{recon}} + w_{kl}L_{kl} + w_pL_p + w_{adv}L_{adv}, \quad (5)$$

with weights $w_{kl} = 10^{-6}$, $w_p = 0.001$, and $w_{adv} = 0.01$. In the second stage, the diffusion UNet is trained to model the conditional distribution of CT given CBCT inputs by progressively denoising latent vectors corrupted with Gaussian noise. The objective function minimizes the mean squared error (MSE) between the predicted noise ϵ_θ and the ground-truth noise ϵ as in (6):

$$L_{\text{diff}} = \|\epsilon - \epsilon_\theta(z_t, t)\|_2^2, \quad (6)$$

where z_t is the nois latent representation at timestep t. This formulation enables the network to learn the iterative denoising process central to diffusion modeling. In our inference, the denoising block in Fig. 1 follows the standard DDPM latent diffusion formulation. Given a clean latent z_0, the forward process adds Gaussian noise using a linear beta schedule $\beta_{\text{start}} = 0.0015$, $\beta_{\text{end}} = 0.0195$:

$$z_t = \sqrt{\overline{\alpha}_t}z_0 + \sqrt{1-\overline{\alpha}_t}\varepsilon, \varepsilon \sim \mathcal{N}(0, I), \quad (7)$$

where $\alpha_t = 1-\beta_t$ and $\overline{\alpha}_t = \prod_{s=1}^{t} \alpha_t$. The diffusion UNet ε_θ is trained to predict ε from (z_t, t) via the MSE loss in Eq. (6). During sampling, we start from $z_T \sim \mathcal{N}(0,I)$ and iteratively apply the reverse update

$$z_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(z_t - \frac{1-\alpha_t}{\sqrt{1-\alpha_t}}\varepsilon_\theta(z_t, t)\right), \quad (8)$$

This reverse update maps the noisy latent z_t at timestep t to a less noisy latent z_{t-1}, using the predicted noise $\varepsilon_\theta(z_t, t, c)$ and the same diffusion schedule $\{\alpha_t, \alpha_t\}$ as in Eq. (7), and is iterated from $t = T$ down to $t = 1$ to obtain the clean latent z_0. The training stages are optimized with the Adam optimizer at a learning rate of 1×10^{-4}. This two-stage training paradigm ensures that the autoencoder produces robust latent embedding, while the diffusion UNet effectively learns to generate anatomically accurate and HU-consistent sCT volumes from noisy latent inputs.

D. IMPLEMENTATION PLATFORM

The proposed 3D-LDFM was implemented using the The proposed 3DLDFM was implemented in Python using the PyTorch deep learning framework, with MONAI providing utilities for volumetric medical image processing. Training and inference were performed on NVIDIA A100 GPUs with 40 GB of VRAM. To efficiently handle the memory demands of full 3D volumes, we enabled automatic mixed precision (AMP), which reduced GPU memory usage and accelerated training without degrading numerical stability.

E. EVALUATION STRATEGY AND METRICS

To rigorously evaluate the effectiveness of the proposed 3D latent diffusion model (3D-LDM) for CBCT-to-CT synthesis, we conducted a comprehensive quantitative assessment of the generated synthetic CT (sCT) volumes against ground-truth planning CTs. Following the evaluation protocol established in the SynthRAD2023 challenge, all similarity metrics were computed within the dilated body contour masks $\mathcal{B}$, ensuring that performance was measured in clinically relevant anatomical regions while excluding background noise.

Three complementary metrics were employed to characterize image fidelity: Mean Absolute Error (MAE), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index Measure (SSIM). Together, these metrics capture voxel-level intensity accuracy, noise robustness, and perceptual/structural fidelity. The masked MAE quantifies the average voxel-wise absolute difference between the generated sCT and the reference CT, normalized by the number of voxels within the anatomical mask as in (9):

$$\text{MAE(CT, sCT)} = \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}|\text{CT}_i - \text{sCT}_i|. \quad (9)$$

Lower MAE values indicate closer correspondence in Hounsfield units (HU), a critical factor for dose calculation in adaptive radiotherapy. To assess the ability of the model to preserve signal quality relative to image noise, we computed

the masked PSNR, defined as in (10):

$$\mathrm{PSNR}\,(\mathrm{CT},\,\mathrm{sCT}) = 10\log_{10}\left(\frac{Q^2}{\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}(\mathrm{CT}_i-\mathrm{sCT}_i)^2}\right), \tag{10}$$

where Q denotes the dynamic range of voxel intensities, clipped to $[-1024, 3000]$ HU in our experiments. A higher PSNR value reflects improved preservation of intensity contrast and lower noise variance in the synthesized images.

Finally, to evaluate perceptual similarity and structural integrity, we used the masked SSIM, which jointly considers luminance, contrast, and structural information between the sCT and CT. For each voxel i, SSIM is computed over a local $7\times7\times7$ neighborhood as in (11):

$$\mathrm{SSIM}_i\,(\mathrm{CT},\,\mathrm{sCT}) = \frac{\left(2\mu_{\mathrm{CT}_i}\mu_{\mathrm{sCT}}+c_1\right)\left(2\sigma_{\mathrm{CT,sCT}}+c_2\right)}{\left(\mu^2_{\mathrm{CT}_i}+\mu^2_{\mathrm{sCT}_i}+c_1\right)\left(\sigma^2_{\mathrm{CT}_i}+\sigma^2_{\mathrm{sCT}_i}+c_2\right)}, \tag{11}$$

where μ and σ are the local means and standard deviations, $\sigma_{\mathrm{CT,sCT}}$ denotes the local covariance, and the constants are defined as $c_1 = (0.01L)^2$, $c_2 = (0.03L)^2$, with L representing the intensity dynamic range. The final SSIM score is obtained by averaging over all voxels within the mask:

$$\mathrm{SSIM}(\mathrm{CT},\,\mathrm{sCT}) = \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\mathrm{SSIM}_i\,(\mathrm{CT},\,\mathrm{sCT}) \tag{12}$$

This metric emphasizes structural preservation and is particularly sensitive to blurring, misalignment, and contrast inconsistencies-factors that strongly impact organ delineation and adaptive treatment planning.

To sum up, this multi-metric evaluation protocol provides a robust and multifaceted assessment of our model's performance. By jointly analyzing voxel-level HU fidelity, intensity quality relative to noise, and perceptual structural consistency, we ensure that the proposed 3D-LDM is validated not only in terms of numerical accuracy but also with respect to clinically meaningful imaging quality essential for safe and effective IGART.

F. BENCHMARK MODELS

To establish a rigorous comparative baseline, our proposed 3D Latent Diffusion Model was benchmarked against four widely recognized state-of-the-art approaches representing both transformer-based and generative approaches namely Swin Transformer with UNet Regularization (Swin UNETR), nnU-Net (no-new-Net), Cycle-Consistent Generative Adversarial Network (CycleGAN), and Conditional Generative Adversarial Network (Pix2PixGAN).

Swin UNETR integrates the hierarchical Swin Transformer as an encoder with a UNet-like decoder connected via skip connections [37], [38]. This architecture has been shown to capture long-range dependencies and multi-scale contextual information in volumetric medical imaging tasks, making it a strong candidate for CBCT-to-CT synthesis where global anatomical consistency is critical.

nnU-Net [39] is a self-configuring deep learning framework that automatically adapts preprocessing, architecture, and training schedules to the dataset at hand. It has consistently served as a robust baseline across numerous medical imaging challenges, demonstrating strong generalization without extensive manual tuning [40]. For CBCT-to-CT translation, nnU-Net provides a powerful CNN-based comparator that establishes the effectiveness of automated, task-adaptive network design.

CycleGAN enables unpaired image-to-image translation by enforcing forward and backward cycle-consistency constraints between two domains [41]. This property has made CycleGAN widely adopted in medical image synthesis scenarios, particularly where paired CBCT-CT datasets are limited or imperfectly aligned [20], [42]. Its inclusion as a benchmark highlights the relative advantage of diffusion-based approaches in handling domain shifts and artifact-rich CBCT data.

Pix2PixGAN directly learns a mapping between paired input and target images under supervised conditions. While effective in scenarios with well-curated paired datasets, Pix2Pix is prone to over-smoothing and can be sensitive to patient misalignments. Nevertheless, it remains a widely recognized baseline for supervised synthesis tasks in radiotherapy imaging [43]. Together, these four baselines span transformer-driven architectures, adaptive CNN frameworks, and GAN-based paired/unpaired synthesis methods. Benchmarking against them provides a comprehensive and fair assessment of our proposed diffusion-based approach, ensuring that improvements in HU fidelity, structural preservation, and robustness are established relative to the best available alternatives.

IV. RESULTS

We comprehensively evaluated the proposed 3D-LDM against state-of-the-art baselines across multiple anatomical sites, using standardized metrics (MAE, PSNR, and SSIM) on the SynthRAD2023 dataset. The results demonstrate consistent improvements in voxel-level fidelity, noise robustness, and structural preservation, validating the clinical readiness of our approach for CBCT-to-CT synthesis in adaptive radiotherapy workflows.

A. QUANTITATIVE RESULTS AND BENCHMARKING

Figure 3–5 summarize the quantitative performance of 3D-LDM compared to SwinUNeTr, nnUNet, CycleGAN, and Pix2Pix across head-and-neck (HN), thorax (TH), and abdominal (AB) sites. For PSNR (**Figure 3**), 3D-LDM consistently achieved higher values, indicating improved signal-to-noise ratios and enhanced robustness against CBCT-specific artifacts such as scatter and streaking. Notably, these performance gains were sustained across thoracic and abdominal regions, underscoring the generalizability of the diffusion framework. With respect to SSIM

(**Figure 4**), the results confirm that 3D-LDM preserves fine structural and contrast details more effectively than competing models. Elevated SSIM scores, particularly in the head-and-neck and thorax datasets, demonstrate the ability of latent diffusion to balance global anatomical consistency with local texture fidelity, a critical requirement for organ delineation and dose recalculation in IGART. Collectively, these findings establish 3D-LDM as a new benchmark in quantitative performance, achieving reduced noise propagation and superior structural similarity compared with leading CNN and GAN architectures. Finally, for MAE (**Figure 5**), the proposed 3D-LDM attained the lowest error distributions across all anatomical regions, reflecting superior HU fidelity relative to CNN- and GAN-based approaches. This advantage was most pronounced in the head-and-neck cohort.

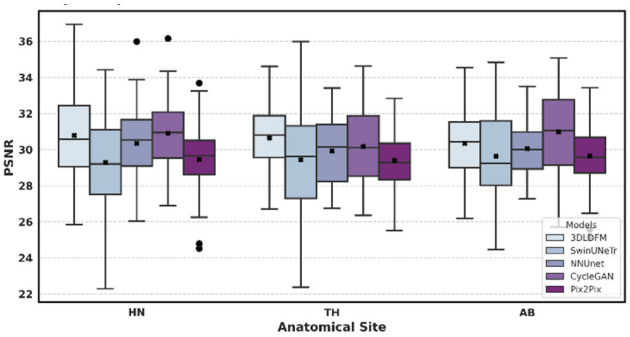

FIGURE 4. Boxplot comparison of SSIM across anatomical sites. The proposed 3D-LDM achieves consistently higher SSIM scores, particularly in HN and TH, highlighting its ability to preserve anatomical structure and contrast detail essential for adaptive radiotherapy.

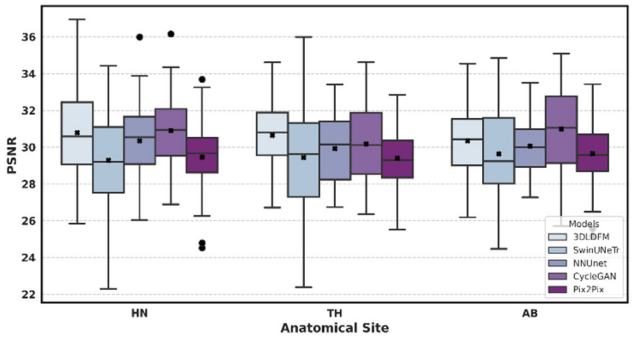

FIGURE 3. Boxplot comparison of PSNR across anatomical sites. 3D-LDM demonstrates higher PSNR values across HN, TH, and AB, indicating improved robustness to noise and preservation of signal intensity compared to baseline methods.

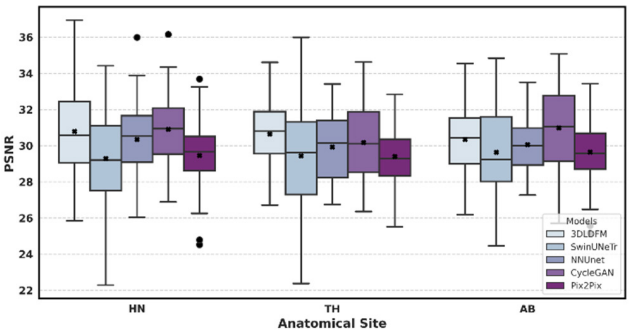

FIGURE 5. Boxplot comparison of MAE across HN, TH, and AB sites. The proposed 3D-LDM consistently achieves the lowest error distributions.

B. COMPARATIVE ANALYSIS WITH STATE-OF-THE-ART

To further validate the effectiveness of our proposed 3DLDFM, we compared its performance with leading state-of-the-art baselines, including SwinUNeTr, nnUNet, CycleGAN, and Pix2Pix. Quantitative results across anatomical regions are summarized in Tables 3–5, reporting mean absolute error, peak signal-to-noise ratio, and structural similarity index measure.

MAE results (**Table 3**) demonstrate that 3DLDFM achieves the lowest voxel-wise error across all anatomical sites, with an overall average of 51.40 ± 11.91, compared to 66.91 ± 19.68 for SwinUNeTr and 53.11 ± 13.04 for CycleGAN. This consistent reduction in error highlights the superior HU fidelity of diffusion-based synthesis, a critical factor for accurate dose calculation.

For PSNR (**Table 4**), 3DLDFM achieved the highest overall score (30.60 ± 2.00), outperforming both CNN and GAN models. The improvement was especially pronounced in head-and-neck and abdominal cohorts, where CBCT images typically suffer from high levels of scatter and streaking artifacts. This finding underscores the robustness of the latent diffusion framework in preserving signal intensity while mitigating CBCT-specific noise.

TABLE 3. Benchmarking based on MAE across anatomical regions. lower values indicate better HU.

Region	SwinUNetR	NNUNet	CycleGAN	Pix2Pix	3D LDM (Ours)
AB	70.56±19.87	57.12±10.46	54.12±11.66	62.35±12.78	51.66±10.97
HN	62.90±20.27	54.48±12.52	52.81±13.59	58.47±10.60	50.83±11.48
TH	67.45±18.42	55.85±10.54	52.45±11.97	59.63±13.78	51.73±13.30
All	66.91±19.69	55.80±11.22	53.11±13.04	60.12±12.49	51.40±11.91

TABLE 4. Benchmarking based on PSNR across anatomical regions. Higher values indicate superior image quality.

Region	SwinUNetR	NNUNet	CycleGAN	Pix2Pix	3D LDM (Ours)
AB	29.64±2.71	30.06±1.40	30.99±2.21	29.65±1.86	30.35±1.84
HN	29.29±2.64	30.35±1.96	30.90±1.95	29.46±1.80	30.79±2.38
TH	29.44±2.90	29.93±1.86	30.18±2.08	29.40±1.66	30.65±1.72
All	29.45±2.74	30.11±1.76	30.69±2.10	29.50±1.77	30.60±2.00

TABLE 5. Benchmarking based on SSIM across anatomical regions. Higher values indicate improved structural fidelity.

Region	SwinUNetR	NNUNet	CycleGAN	Pix2Pix	3D LDM (Ours)
AB	0.8817±0.0353	0.9040±0.0298	0.9024±0.0408	0.8818±0.0392	0.9103±0.0332
HN	0.8791±0.0398	0.9015±0.0373	0.8987±0.0352	0.8772±0.0349	0.9089±0.0340
TH	0.8838±0.0395	0.8990±0.0304	0.9060±0.0356	0.8795±0.0369	0.9181±0.0338
All	0.8815±0.0382	0.9015±0.0327	0.9024±0.0372	0.8795±0.0369	0.9124±0.0338

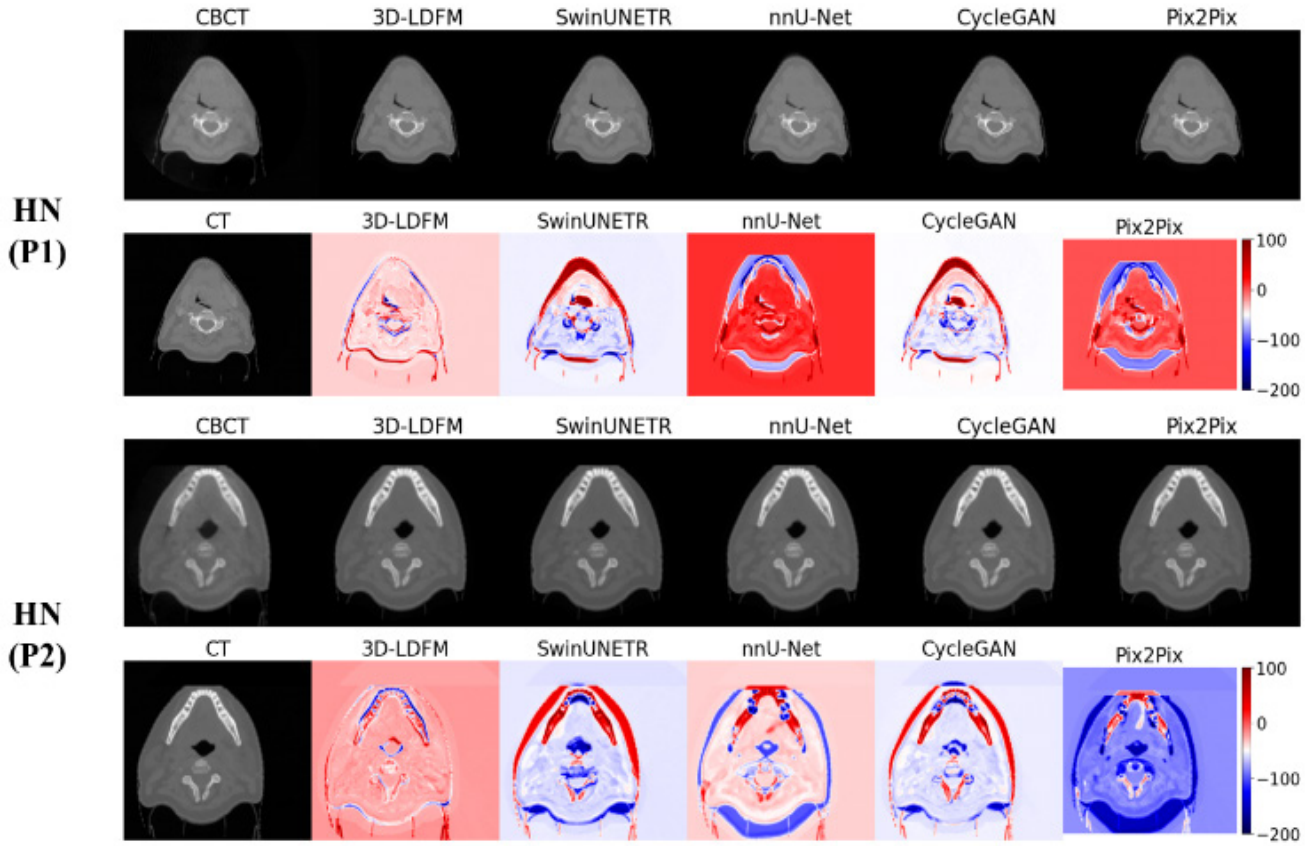

FIGURE 6. Qualitative comparison of sCT generation for head-and-neck cases. Top rows show CBCT inputs and corresponding sCT reconstructions across models; bottom rows depict error maps against ground truth CT. 3D-LDFM achieves sharper reconstructions with reduced HU deviations compared to SwinUNETR, nnU-Net, CycleGAN, and Pix2Pix.

Finally, SSIM results (**Table 5**) confirm that 3DLDFM provides the best structural fidelity, with an average score of 0.912 ± 0.033, surpassing nnUNet (0.902 ± 0.032) and CycleGAN (0.902 ± 0.037). Notably, 3DLDFM maintained high SSIM across all anatomical regions, with the highest structural preservation observed in the thoracic dataset

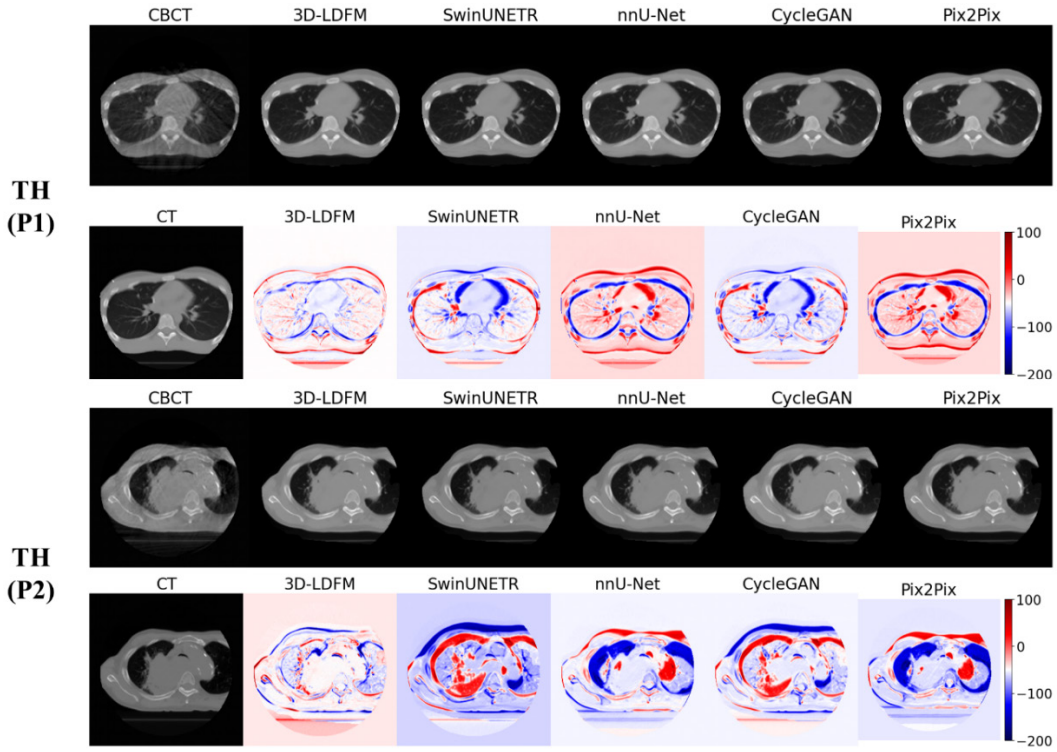

FIGURE 7. Thoracic sCT synthesis results. The proposed 3D-LDFM effectively reduces scatter and streak artifacts, yielding improved fidelity in lung and mediastinal regions relative to CNN- and GAN-based baselines. Error maps confirm lower voxel-wise deviations and superior HU accuracy.

(0.918 ± 0.034). This demonstrates the model's ability to balance global anatomical accuracy with local texture detail, an essential requirement for contour propagation and adaptive radiotherapy planning. Collectively, these results establish 3DLDFM as a new benchmark for CBCT-to-CT synthesis, outperforming both CNN and GAN architectures in terms of HU accuracy, noise robustness, and structural similarity.

C. PERFORMANCE ACROSS ANATOMICAL SITES

To further elucidate the robustness of the proposed 3D-LDFM across different anatomical regions, we conducted a qualitative comparison with state-of-the-art architectures including SwinUNETR, nnU-Net, CycleGAN, and Pix2Pix. This section analyzes the performance of the 3D LDM against benchmark models using visual comparisons of sCT volumes and their corresponding absolute error maps (|sCT−CT|) across the Head and Neck (HN), Thorax (TH), and Abdomen (AB) regions. The error maps are displayed on a window of [−200,200] HU, with colors indicating the magnitude of the HU deviation from the ground truth CT. Across all anatomical sites, 3D-LDFM demonstrates superior capacity to suppress CBCT-specific artifacts, preserve fine structural details, and maintain Hounsfield Unit accuracy, underscoring its clinical viability for image-guided adaptive radiotherapy.

1) HEAD AND NECK SYNTHESIS

Figure 6 illustrates the performance comparison for two representative HN patients (P1 and P2). The HN region is anatomically complex, featuring small, high-density structures (bone, dental fillings) adjacent to low-density air cavities, which typically induces significant streaking artifacts in CBCT (P2 CBCT). In the error maps, the U-Net and GAN-based models (SwinUNetR, nNU-Net, CycleGAN, Pix2Pix) exhibit widespread large HU deviations (deep red and blue) particularly at the air-tissue interfaces and within bony structures (vertebrae, mandible). This indicates an inability to accurately correct partial volume effects and scatter. In contrast, the 3D-LDFM error maps show significantly less intense and more localized errors, primarily restricted to the thin bone cortex or sharp tissue edges. For P2, where the CBCT noise is severe, 3D-LDFM effectively suppresses the widespread noise while maintaining the clear demarcation of the spinal cord and soft tissues, demonstrating its robust noise-to-signal generation capacity derived from the diffusion process.

2) THORAX SYNTHESIS

The Thorax region presents unique challenges due to low-density lung tissue and motion-induced artifacts around the diaphragm and heart. **Figure 7** shows that the benchmark models, particularly the GANs and SwinUNetR, struggle with the low-density environment, displaying large areas of error within the lung fields and surrounding the mediastinum (P1 and P2). The red/blue error accumulation around the ribs and vertebrae in competitor models signifies poor bone-soft tissue boundary reconstruction. Conversely,

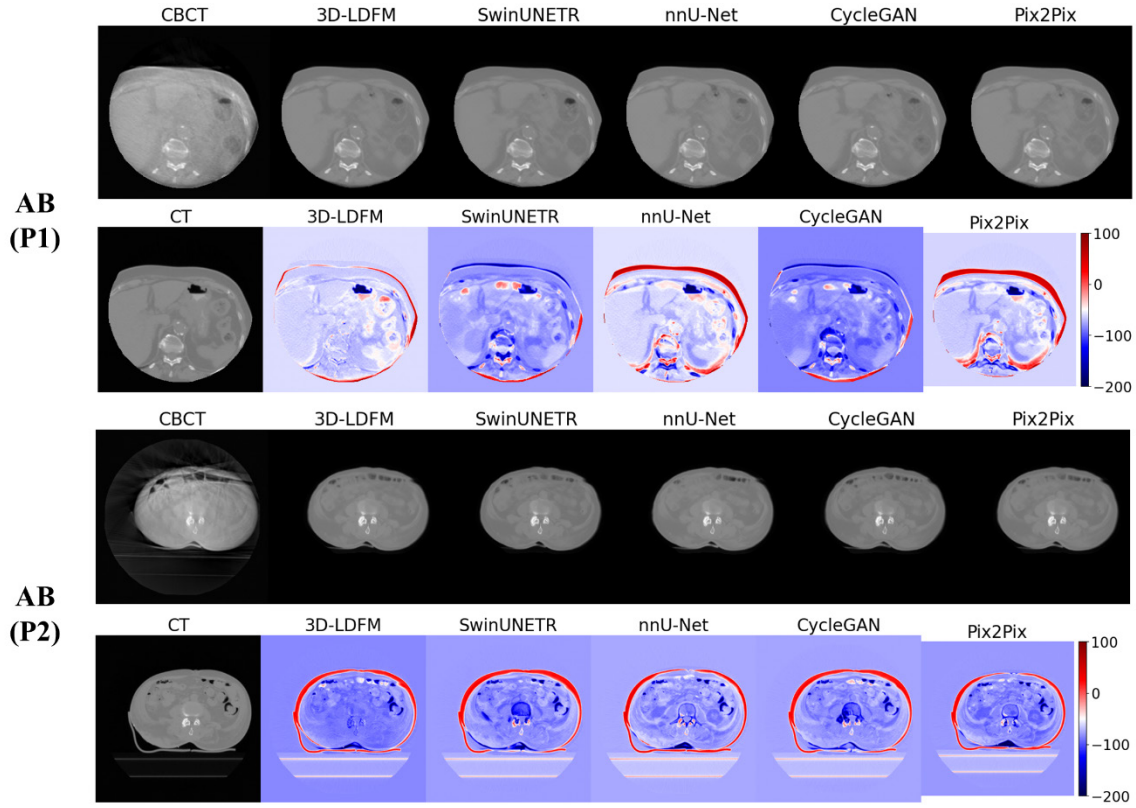

FIGURE 8. Abdominal sCT synthesis results. 3D-LDFM demonstrates enhanced HU fidelity and structural preservation in liver and bowel regions, while baseline models exhibit texture blurring or intensity shifts. Error maps highlight the superior artifact suppression and robustness of the diffusion framework.

the 3D-LDFM consistently exhibits the cleanest error maps in the low-density lung parenchyma, validating the quantitative superiority in SSIM by accurately preserving the complex vascular structures and soft tissue textures in the lung. The 3D-LDFM's error is minimized and confined primarily to the pleural and pericardial interfaces, confirming its architectural advantage in handling low-contrast, heterogeneous regions common in the thoracic cavity.

3) THORAX SYNTHESIS

The Abdomen is challenging due to the large presence of heterogeneous soft tissues, variable fat content, and patient motion leading to truncation and cupping artifacts (P2 CBCT). As depicted in **Figure 8**, the benchmark models show extensive HU errors in the periphery and within deep abdominal organs (liver, kidneys, spleen). The nNU-Net and SwinUNetR demonstrate structural blurring, and the GANs show pronounced boundary errors. The 3D-LDFM sCT, however, visually appears closest to the ground truth CT. The corresponding error maps show significantly less saturated and less widespread errors across both P1 and P2. Notably, for the severe truncation artifacts in P2, the 3D-LDFM successfully provides a much smoother and more accurate HU profile throughout the entire cross-section than the competitors, affirming its ability to generalize and correct complex, non-local CBCT artifacts. This qualitative consistency across all three sites provides robust visual evidence supporting the statistical significance of the quantitative results.

V. DISCUSSION

The findings of this study demonstrate that the proposed 3DLDFM achieves superior performance over established CNN-based and GAN-based methods for CBCT-to-CT synthesis across multiple anatomical regions. By integrating a variational autoencoder for latent space compression with a diffusion-based generative process, the framework consistently delivered lower voxel-wise errors (MAE), higher signal fidelity (PSNR), and improved structural preservation (SSIM). These improvements highlight the advantages of diffusion models in capturing the conditional distribution of medical imaging data, thereby overcoming key limitations of deterministic CNNs and adversarial networks.

A critical strength of 3DLDFM lies in its ability to balance HU accuracy and structural fidelity, both of which have the potential for IGART, it validated in external in large cohort. Lower MAE values translate directly into more accurate HU calibration, which underpins reliable dose recalculation and adaptive treatment planning. At the same time, higher SSIM scores reflect superior structural consistency, ensuring that fine anatomical details such as organ boundaries and

bony landmarks are faithfully preserved. These dual gains are particularly relevant in head-and-neck cases, where complex anatomical variability and the presence of dental artifacts have historically challenged existing synthesis models.

Compared with GAN-based approaches such as CycleGAN and Pix2Pix, which often suffer from training instability and hallucinated features, diffusion models provide a more stable and principled generative framework. The iterative denoising process enables robust reconstruction of clinically meaningful features while mitigating the over-smoothing commonly observed in CNN-based regression models. Furthermore, by operating in a compressed latent space, 3DLDFM achieves computational feasibility for volumetric data without sacrificing accurate essential requirement for time-sensitive online ART workflows [24].

From a clinical standpoint, the proposed model advances the feasibility of simulation-free adaptive workflows, where daily CBCT could serve as the sole imaging modality for replanning [1], [2], [3]. This shift has the potential to reduce patient burden, streamline treatment pipelines, and minimize the dependency on frequent planning CT acquisitions. Importantly, the multi-site evaluation confirms that the model generalizes across diverse anatomical regions and acquisition protocols, suggesting robustness against domain shifts that frequently hinder real-world adoption.

Nevertheless, several limitations warrant discussion. First, while quantitative metrics (MAE, PSNR, SSIM) provide strong evidence of performance, clinical validation through dose recalculation studies and contour propagation accuracy remains necessary to establish direct therapeutic benefits. Second, the current framework relies on paired CBCT-CT training data; extending to semi-supervised or unpaired settings would further improve scalability. Finally, although latent diffusion significantly reduces computational cost relative to pixel-space diffusion, real-time inference speed remains a practical barrier for widespread clinical integration and could benefit from optimized sampling strategies or model distillation.

Looking forward, future work should focus on three directions: (i) integrating clinical endpoints such as dose–volume histogram (DVH) comparisons and γ-index analysis into the evaluation pipeline; (ii) extending the framework to multi-modal synthesis tasks, including MRI-to-CT and CBCT artifact correction; and (iii) developing lightweight, accelerated diffusion architectures suitable for deployment in radiotherapy treatment rooms. Finally, this study establishes 3DLDFM as a state-of-the-art solution for CBCT-to-CT synthesis, combining high HU fidelity, structural accuracy, and cross-site generalizability.

VI. CONCLUSION

In this work, we introduced a 3DLDFM for CBCT-to-CT synthesis, designed to address the persistent challenges of HU inaccuracy, structural degradation, and poor generalization that limit existing CNN- and GAN-based solutions in image-guided adaptive radiotherapy. By combining latent space compression with a diffusion-driven generative process, the framework explicitly models the conditional distribution of CT given CBCT while remaining computationally tractable for full 3D volumes. From an imaging standpoint, the proposed 3DLDFM achieves consistent and substantial gains across multiple, well-established objective metrics. Quantitative evaluations on the SynthRAD2023 dataset demonstrate that 3DLDFM yields the lowest voxel-wise error, the highest peak signal-to-noise ratio, and the best structural similarity index across head-and-neck, thorax, and abdominal cohorts when compared with strong CNN- and GAN-based baselines (SwinUNETR, nnUNet, CycleGAN, Pix2Pix).

These improvements indicate that the model not only corrects HU values more accurately but also better preserves global anatomy and local texture, setting a new benchmark for CBCT-to-CT translation in terms of HU fidelity, noise robustness, and structural preservation. The necessity of such improvements is directly linked to clinical decision-making in IGART. More accurate HU calibration (reflected in lower MAE and higher PSNR) underpins reliable dose recalculation and adaptive plan optimization, while higher SSIM supports faithful reproduction of organ boundaries and bony landmarks, which are crucial for contour propagation and daily plan adaptation. Although the present study evaluates performance using image-based metrics, these quantities are closely related to downstream dosimetric endpoints and provide strong evidence that diffusion-based synthesis can reduce the uncertainty inherent in CBCT-driven workflows. Finally, this study establishes latent diffusion modeling as a promising and practically viable direction for simulation-free, CBCT-driven adaptive radiotherapy. Future work will extend the evaluation to direct clinical endpoints, including DVH analysis, and will explore domain adaptation and accelerated sampling strategies to ensure robust image translation.

REFERENCES

[1] O. M. Dona Lemus, M. Cao, B. Cai, M. Cummings, and D. Zheng, "Adaptive radiotherapy: Next-generation radiotherapy," *Cancers*, vol. 16, no. 6, p. 1206, Mar. 2024.

[2] M. Guberina, N. Guberina, C. Hoffmann, A. Gogishvili, F. Freisleben, A. Herz, J. Hlouschek, T. Gauler, S. Lang, K. Stähr, B. Höing, C. Pöttgen, F. Indenkämpen, A. Santiago, A. Khouya, S. Mattheis, and M. Stuschke, "Prospects for online adaptive radiation therapy (ART) for head and neck cancer," *Radiat. Oncol.*, vol. 19, no. 1, p. 4, Jan. 2024.

[3] G. Kalinauskaite, L. A. Künzel, A. Kluge, K. Rubarth, J. Dannehl, C. Höhne, M. Beck, D. Zips, and C. Senger, "Optimizing workflow for cone beam computed tomography-based online adaptive radiation therapy toward reduced physician involvement," *Adv. Radiat. Oncol.*, vol. 10, no. 10, Oct. 2025, Art. no. 101874.

[4] P. J. van Houdt, Y. Yang, and U. A. van der Heide, "Quantitative magnetic resonance imaging for biological image-guided adaptive radiotherapy," *Frontiers Oncol.*, vol. 10, Jan. 2021, Art. no. 615643.

[5] E. Mastella, F. Calderoni, L. Manco, M. Ferioli, S. Medoro, A. Turra, M. Giganti, and A. Stefanelli, "A systematic review of the role of artificial intelligence in automating computed tomography-based adaptive radiotherapy for head and neck cancer," *Phys. Imag. Radiat. Oncol.*, vol. 33, Jan. 2025, Art. no. 100731.

[6] G. S. Ibbott, "The need for, and implementation of, image guidance in radiation therapy," *Ann. ICRP*, vol. 47, nos. 3–4, pp. 160–176, Oct. 2018.
[7] E. M. C. Huijben et al., "Generating synthetic computed tomography for radiotherapy: SynthRAD2023 challenge report," *Med. image Anal.*, vol. 97, Oct. 2024, Art. no. 103276.
[8] A. Thummerer, E. van der Bijl, A. Galapon, J. J. C. Verhoeff, J. A. Langendijk, S. Both, C. A. T. van den Berg, and M. Maspero, "SynthRAD2023 grand challenge dataset: Generating synthetic CT for radiotherapy," *Med. Phys.*, vol. 50, no. 7, pp. 4664–4674, Jul. 2023.
[9] D. Lustermans, G. P. Fonseca, V. T. Taasti, A. van de Schoot, S. Petit, W. van Elmpt, and F. Verhaegen, "Image quality evaluation of a new high-performance ring-gantry cone-beam computed tomography imager," *Phys. Med. Biol.*, vol. 69, no. 10, May 2024, Art. no. 105018.
[10] M. Tsarenko and L. Kalashnikova, "Optimization of scanning parameters for CT and CBCT: A systematic review," KPI, Kyiv, Ukraine, Tech. Rep., 2025, vol. 1, no. 17.
[11] M. Pepa, S. Taleghani, G. Sellaro, A. Mirandola, F. Colombo, S. Vennarini, M. Ciocca, C. Paganelli, E. Orlandi, G. Baroni, and A. Pella, "Unsupervised deep learning for synthetic CT generation from CBCT images for proton and carbon ion therapy for paediatric patients," *Sensors*, vol. 24, no. 23, p. 7460, Nov. 2024.
[12] A. Altalib, S. McGregor, C. Li, and A. Perelli, "Synthetic CT image generation from CBCT: A systematic review," *IEEE Trans. Radiat. Plasma Med. Sci.*, vol. 9, no. 6, pp. 691–707, Jul. 2025.
[13] B. Rusanov, G. M. Hassan, M. Reynolds, M. Sabet, P. Rowshanfarzad, N. Bucknell, S. Gill, J. Dass, and M. Ebert, "Transformer CycleGAN with uncertainty estimation for CBCT based synthetic CT in adaptive radiotherapy," *Phys. Med. Biol.*, vol. 69, no. 3, Feb. 2024, Art. no. 035014.
[14] M. A. Mahdi, S. Ahamad, S. A. Saad, A. Dafhalla, R. Qureshi, and A. Alqushaibi, "Weighted fusion transformer for dual PET/CT head and neck tumor segmentation," *IEEE Access*, vol. 12, pp. 110905–110919, 2024.
[15] X. Li, R. Yang, M. Li, X. Li, A. J. Lomax, J. M. Buhmann, and Y. Zhang, "Continuous sPatial-temporal deformable image registration (CPT-DIR) for motion modelling in radiotherapy: Beyond classic voxel-based methods," 2024, *arXiv:2405.00430*.
[16] B. Rigaud, A. Simon, J. Castelli, C. Lafond, O. Acosta, P. Haigron, G. Cazoulat, and R. de Crevoisier, "Deformable image registration for radiation therapy: Principle, methods, applications and evaluation," *Acta Oncologica*, vol. 58, no. 9, pp. 1225–1237, Sep. 2019.
[17] Y. Hu, H. Zhou, N. Cao, C. Li, and C. Hu, "Synthetic CT generation based on CBCT using improved vision transformer CycleGAN," *Sci. Rep.*, vol. 14, no. 1, p. 11455, May 2024.
[18] S. Chen, A. Qin, D. Zhou, and D. Yan, "Technical note: U-Net-generated synthetic CT images for magnetic resonance imaging-only prostate intensity-modulated radiation therapy treatment planning," *Med. Phys.*, vol. 45, no. 12, pp. 5659–5665, Dec. 2018.
[19] L. Deng, J. Hu, J. Wang, S. Huang, and X. Yang, "Synthetic CT generation based on CBCT using respath-cycleGAN," *Med. Phys.*, vol. 49, no. 8, pp. 5317–5329, Aug. 2022.
[20] Y. Liu, Y. Lei, T. Wang, Y. Fu, X. Tang, W. J. Curran, T. Liu, P. Patel, and X. Yang, "CBCT-based synthetic CT generation using deep-attention cycleGAN for pancreatic adaptive radiotherapy," *Med. Phys.*, vol. 47, no. 6, pp. 2472–2483, Jun. 2020.
[21] Y. Zhang, N. Yue, M.-Y. Su, B. Liu, Y. Ding, Y. Zhou, H. Wang, Y. Kuang, and K. Nie, "Improving CBCT quality to CT level using deep learning with generative adversarial network," *Med. Phys.*, vol. 48, no. 6, pp. 2816–2826, Jun. 2021.
[22] B. Rusanov, M. A. Ebert, G. Mukwada, G. M. Hassan, and M. Sabet, "A convolutional neural network for estimating cone-beam CT intensity deviations from virtual CT projections," *Phys. Med. Biol.*, vol. 66, no. 21, Nov. 2021, Art. no. 215007.
[23] J. Zhu, W. Chen, H. Sun, S. Zhi, J. Qin, J. Cai, and G. Ren, "Feature-oriented deep learning framework for pulmonary cone-beam CT (CBCT) enhancement with multi-task customized perceptual loss," 2023, *arXiv:2311.00412*.
[24] A. Altalib, C. Li, and A. Perelli, "Conditional diffusion models for CT image synthesis from CBCT: A systematic review," 2025, *arXiv:2509.17790*.
[25] E. Zhu, A. Muneer, J. Zhang, Y. Xia, X. Li, C. Zhou, J. V. Heymach, J. Wu, and X. Le, "Progress and challenges of artificial intelligence in lung cancer clinical translation," *npj Precis. Oncol.*, vol. 9, no. 1, p. 210, Jul. 2025.
[26] Y. C. I. Chan, M. Li, A. Thummerer, K. Parodi, C. Belka, C. Kurz, and G. Landry, "Minimum imaging dose for deep learning-based pelvic synthetic computed tomography generation from cone beam images," *Phys. Imag. Radiat. Oncol.*, vol. 30, Apr. 2024, Art. no. 100569.
[27] C. Khamfongkhruea, T. Prakarnpilas, S. Thongsawad, A. Deeharing, T. Chanpanya, T. Mundee, P. Suwanbut, and K. Nimjaroen, "Supervised deep learning-based synthetic computed tomography from kilovoltage cone-beam computed tomography images for adaptive radiation therapy in head and neck cancer," *Radiat. Oncol. J.*, vol. 42, no. 3, pp. 181–191, Sep. 2024.
[28] X. Xue, Y. Ding, J. Shi, X. Hao, X. Li, D. Li, Y. Wu, H. An, M. Jiang, W. Wei, and X. Wang, "Cone beam CT (CBCT) based synthetic CT generation using deep learning methods for dose calculation of nasopharyngeal carcinoma radiotherapy," *Technol. Cancer Res. Treatment*, vol. 20, Jan. 2021, Art. no. 15330338211062415.
[29] Y. Koike, H. Takegawa, Y. Anetai, S. Nakamura, K. Yoshida, A. Yoshida, M. Yui, K. Hirota, K. Ueda, and N. Tanigawa, "Cone-beam CT to CT image translation using a transformer-based deep learning model for prostate cancer adaptive radiotherapy," *J. Imag. Informat. Med.*, vol. 38, no. 4, pp. 2490–2499, Nov. 2024.
[30] Y. Hu, M. Cheng, H. Wei, and Z. Liang, "A joint learning framework for multisite CBCT-to-CT translation using a hybrid CNN-transformer synthesizer and a registration network," *Frontiers Oncol.*, vol. 14, Aug. 2024, Art. no. 1440944.
[31] J. Peng, R. L. J. Qiu, J. F. Wynne, C.-W. Chang, S. Pan, T. Wang, J. Roper, T. Liu, P. R. Patel, D. S. Yu, and X. Yang, "CBCT-based synthetic CT image generation using conditional denoising diffusion probabilistic model," *Med. Phys.*, vol. 51, no. 3, pp. 1847–1859, Mar. 2024.
[32] X. Chen, R. L. J. Qiu, J. Peng, J. W. Shelton, C.-W. Chang, X. Yang, and A. H. Kesarwala, "CBCT-based synthetic CT image generation using a diffusion model for CBCT-guided lung radiotherapy," *Med. Phys.*, vol. 51, no. 11, pp. 8168–8178, Nov. 2024.
[33] Y. Zhang, L. Li, J. Wang, X. Yang, H. Zhou, J. He, Y. Xie, Y. Jiang, W. Sun, X. Zhang, G. Zhou, and Z. Zhang, "Texture-preserving diffusion model for CBCT-to-CT synthesis," *Med. Image Anal.*, vol. 99, Jan. 2025, Art. no. 103362.
[34] C. Hu, N. Cao, X. Li, Y. He, and H. Zhou, "CBCT-to-CT synthesis using a hybrid U-Net diffusion model based on transformers and information bottleneck theory," *Sci. Rep.*, vol. 15, no. 1, p. 10816, Mar. 2025.
[35] J. Xie, H.-C. Shao, Y. Li, and Y. Zhang, "Prior frequency guided diffusion model for limited angle (LA)-CBCT reconstruction," *Phys. Med. Biol.*, vol. 69, no. 13, Jul. 2024, Art. no. 135008.
[36] M. A. Mahdi, M. Al-Shalabi, E. T. Alnfrawy, R. Elbarougy, M. U. Hadi, and R. F. Ali, "3D latent diffusion model for MR-only radiotherapy: Accurate and consistent synthetic CT generation," *Diagnostics*, vol. 15, no. 23, p. 3010, Nov. 2025.
[37] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, "Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images," in *Proc. Int. MICCAI Brainlesion Workshop*, 2022, pp. 272–284.
[38] M. A. Mahdi, S. Ahamad, S. A. Saad, A. Dafhalla, A. Alqushaibi, and R. Qureshi, "Segmentation of head and neck tumors using dual PET/CT imaging: Comparative analysis of 2D, 2.5D, and 3D approaches using UNet transformer," *Comput. Model. Eng. Sci.*, vol. 141, no. 3, pp. 2351–2373, 2024.
[39] F. Isensee, J. Petersen, A. Klein, D. Zimmerer, P. F. Jaeger, S. Kohl, J. Wasserthal, G. Koehler, T. Norajitra, S. Wirkert, and K. H. Maier-Hein, "nnU-Net: Self-adapting framework for U-Net-based medical image segmentation," 2018, *arXiv:1809.10486*.
[40] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, Feb. 2021.
[41] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2242–2251.
[42] L. Gao, K. Xie, X. Wu, Z. Lu, C. Li, J. Sun, T. Lin, J. Sui, and X. Ni, "Generating synthetic CT from low-dose cone-beam CT by using generative adversarial networks for adaptive radiotherapy," *Radiat. Oncol.*, vol. 16, no. 1, p. 202, Dec. 2021.
[43] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5967–5976.

MOHAMMED AL-SHALABI received the master's degree in computer science from Yarmouk University, Irbid, Jordan, in 2005, and the Ph.D. degree in computer science, specializing in wireless sensor networks from USM, Malaysia, in 2018. He was a Lecturer in computer science with Philadelphia University, from 2005 to 2007. After that, he was a Lecturer of computer science with Sharjah University, United Arab Emirates, from 2007 to 2017. He is currently an Assistant Professor of computer science with the College of Computer Science and Engineering, University of Hail, Saudi Arabia. He has many publications in different journals, in addition to attending many conferences. His research interests include artificial intelligence, wireless sensor networks, data security, and the IoT.

MOHAMMED A. MAHDI received the Ph.D. degree in computer science from the School of Computer Science, University Science Malaysia (USM), in 2016. He is currently an Associate Professor of computer science with the College of Computer Science and Engineering, University of Hail, Saudi Arabia. He is an Expert in the field of routing protocols and mobile ad-hoc networks with ten years of research and academic experience. He has contributed to publishing more than 25 research articles in journals and conference proceedings. He is participating in academic and institutional development activities at the college and university levels. His research interests include artificial intelligence, medical imaging, wireless networks, mobile networks, cryptography and network security, artificial intelligence, the IoT, and cloud computing.

REDA ELBAROUGY received the B.Sc. and M.Sc. degrees in computer science from Mansoura University, Egypt, in May 1997 and February 2006, respectively. He is currently an Assistant Professor with the Department of Artificial Intelligence and Data Science, College of Computer Science and Engineering, University of Hail, Saudi Arabia. He was with the Faculty of Science, Mansoura University, from 1999 to 2009. In July 2009, he joined Japan Advanced Institute of Science and Technology (JAIST), Japan, as a Ph.D. Student. From September 2014 to August 2019, he was with the Mathematics Department, Faculty of Science, Damietta University as an Assistant Professor. In 2017, he was a Postdoctoral Researcher funded by JSPS to conduct research at Japan Advanced Institute of Science and Technology (JAIST), from June 2017 to April 2019. He has published several journals and at conference papers in these areas. He is looking into applying machine learning and deep learning techniques to automate malware analysis and detection. His research interests include machine learning, artificial intelligence, natural language processing, Arabic text summarization, speech analysis, and speech emotion recognition.

EHAB TAWFEEK ALNFRAWY received the Ph.D. degree in computer science from Cairo University, Egypt, in 2010, specializing in cryptography and security. He is currently an Assistant Professor with the Information Security Department, College of Computer Science and Engineering, University of Hail, Saudi Arabia. With more than 28 years of academic experience, he has held positions at institutions, such as Al-Azhar University and Najran University, Saudi Arabia, and Sina University, Egypt. He has published numerous articles in these fields and actively contributes to advancing knowledge in cybersecurity and cryptography. His research interests include cybersecurity, the IoT security, and computer forensics, with a focus on enhancing security protocols through innovative methodologies.

MUHAMMAD USMAN HADI (Member, IEEE) received the M.S. and Ph.D. degrees from the University of Bologna, Italy. He was a Postdoctoral Researcher with Aalborg University, Denmark, in close collaboration with Nokia. He was a Visiting Researcher with ESIEE Paris, France, and Nokia Bell Laboratories. He is currently an Assistant Professor with the School of Engineering, Ulster University, Belfast, U.K. He has authored more than 60 journal articles and transactions. As a Principal Investigator, he has received EPSRC DTNET+, Innovate U.K., British Council Going Global Partnerships, Garfield Weston Trust, R&I funds, and EPSRC, DFE, and ISPF research grants as a Co-Investigator. He has been included in the top 2% of researchers for four consecutive years, since 2021. His research interests include machine learning for engineering applications, autonomous aerial vehicles, fiber wireless communications, microwave photonics, and devices for telecommunications.

RAO FAIZAN ALI (Member, IEEE) received the bachelor's degree in computer science from COMSATS University Islamabad, Pakistan, the M.Phil. degree in computer science from the University of Management and Technology, Lahore, Pakistan, and the Ph.D. degree from University Technology PETRONAS, Malaysia. He has ten years of experience in teaching and research. He is currently a Lecturer with the School of Computing, University of Kent, U.K. He has been a consultant with various computer science positions in financial, consulting, academia, and the government sector. His research interests include generative artificial intelligence, bioinformatics, AI and data science, information security, and behavioral information.