Medical Image Analysis 99 (2025) 103362

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Texture-preserving diffusion model for CBCT-to-CT synthesis

Youjian Zhang [a], Li Li [a], Jie Wang [a], Xinquan Yang [a], Haotian Zhou [a], Jiahui He [b,e], Yaoqin Xie [b], Yuming Jiang [c], Wei Sun [d], Xinyuan Zhang [f], Guanqun Zhou [a,b], Zhicheng Zhang [a,b,*]

[a] JancsiLab, JancsiTech, Hong Kong, China
[b] Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Guangdong, 518055, China
[c] Department of Radiation Oncology, Wake Forest University School of Medicine, Winston Salem, NC, USA
[d] University of Science and Technology of China, Anhui, 230026, China
[e] School of Computer Science, Faculty of Science and Engineering, University of Nottingham Ningbo China, Zhejiang 315100, China
[f] School of Biomedical Engineering, Southern Medical University, Guangdong, China

ARTICLE INFO

Keywords:
Cone-beam CT
CBCT-to-CT synthesis
Diffusion model

ABSTRACT

Cone beam computed tomography (CBCT) serves as a vital imaging modality in diverse clinical applications, but is constrained by inherent limitations such as reduced image quality and increased noise. In contrast, computed tomography (CT) offers superior resolution and tissue contrast. Bridging the gap between these modalities through CBCT-to-CT synthesis becomes imperative. Deep learning techniques have enhanced this synthesis, yet challenges with generative adversarial networks persist. Denoising Diffusion Probabilistic Models have emerged as a promising alternative in image synthesis. In this study, we propose a novel texture-preserving diffusion model for CBCT-to-CT synthesis that incorporates adaptive high-frequency optimization and a dual-mode feature fusion module. Our method aims to enhance high-frequency details, effectively fuse cross-modality features, and preserve fine image structures. Extensive validation demonstrates superior performance over existing methods, showcasing better generalization. The proposed model offers a transformative pathway to augment diagnostic accuracy and refine treatment planning across various clinical settings. This work represents a pivotal step toward non-invasive, safer, and high-quality CBCT-to-CT synthesis, advancing personalized diagnostic imaging practices.

1. Introduction

Cone-beam computed tomography (CBCT), with its lower radiation doses and aptness for localized anatomical evaluations, stands as a fundamental tool across diverse clinical scenarios, enabling real-time visualization and precise guidance for intricate procedures such as image-guided surgeries (Siewerdsen, 2011; Ujiie et al., 2017), radiation therapy (Barney et al., 2011; Mututantri-Bastiyange and Chow, 2020), and dental interventions. However, inherent limitations in CBCT systems, including compromised image quality, increased noise levels, and the presence of artifacts (Patel et al., 2015; Liang et al., 2019b), restrict its adaptability within specific clinical contexts. In contrast, computed tomography (CT) technology, which offers superior resolution, fewer artifacts, and improved soft tissue contrast, is crucial for precise anatomical delineation (Li et al., 2023b) and comprehensive pathological evaluations (Jiang et al., 2023). However, the different imaging characteristics of CBCT and CT necessitate an intermediary mechanism to bridge this disparity, resulting in the requirement of CBCT-to-CT synthesis.

Advancements in deep learning (DL)-based X-ray enhancement techniques have significantly facilitated CBCT-to-CT synthesis in both the image and projection domains. Previous research used sophisticated architectures such as Pix2Pix (Jiang et al., 2022), CycleGAN (Liang et al., 2019a; Liu et al., 2020) to generate synthetic CT (sCT) images from CBCT data, effectively addressing inherent image artifacts and reducing dependence on paired datasets. These models can effectively learn the relationship between the two modalities and generate synthetic CT images with improved quality and structural similarity to conventional CT scans. The advantage of deep learning in CBCT-to-CT synthesis lies in its ability to extract intricate features and patterns from CBCT data, enabling more accurate image synthesis. However, challenges persist in Generative Adversarial Networks (GANs) in CBCT-to-CT synthesis, including issues in network training and application in medical imaging, leading to the potential introduction of unrealistic structures and complex training complexities (Li et al., 2023a; Wang et al., 2020).

* Corresponding author at: JancsiLab, JancsiTech, Hong Kong, China.
E-mail addresses: gqzhou@stanford.edu (G. Zhou), zhangzhicheng13@mails.ucas.edu.cn (Z. Zhang).

https://doi.org/10.1016/j.media.2024.103362
Received 19 December 2023; Received in revised form 7 August 2024; Accepted 29 September 2024
Available online 9 October 2024

Recent years have witnessed the emergence of Denoising Diffusion Probabilistic Models (DDPMs) as promising tools for diverse domain image synthesis (Ho et al., 2020; Xia et al., 2023; Su et al., 2022; Parmar et al., 2023). DDPMs involve stochastic processes that encompass both forward and reverse diffusion, allowing a comprehensive exploration of natural image distributions. Although the original DDPMs function as unsupervised learning algorithms, producing images reflecting a random distribution within the target domain, the challenge in CBCT-to-CT synthesis lies in its conditional nature, requiring synthesizing sCT images to correspond specifically to CBCT images, rather than generating random CT images. Addressing this issue led to the introduction of the Conditional Denoising Diffusion Probabilistic Model (CDDPM) which adopts a temporal embedding U-Net structure with the combination of residual and attention mechanism blocks. Eventually, CDDPM transforms standard Gaussian noise into the desired CT distribution conditioned on CBCT. Additionally, CDDPM has been used to explore various conditional noise application methods to enhance synthesis (Peng et al., 2023). Another form of conditional diffusion model is the Denoising Diffusion Medical Models (DDMM). These models feature multi-task branch structures, sharing noise schedules and latent spaces, ensuring semantic consistency. DDMM enables the simultaneous generation of images from the target domain and additional tasks such as segmentation (Huy and Quan, 2023). With the assistance of GAN, DDPM-based methods can generate sCT from CBCT in an unsupervised manner. For instanse, SynDiff (Özbey et al., 2023) combines adversarial diffusion models with unsupervised training, using adversarial projection to refine image sampling accuracy within a cycle-consistent framework. Moreover, FGDM (Li et al., 2023a) guides image diffusion and denoising through frequency domain analysis, combining low- and high-pass filtering to preserve and restore mid- and high-frequency image details.

Despite the achievements of these diffusion models in certain aspects, there remain shortcomings in generating detailed images of the target domain, particularly in high-frequency information. Additionally, limited by insufficient training data, the generative capabilities of these models might be inadequate, potentially leading to a loss of image details and consequently affecting the integrity of the generated images. To address these challenges, our work introduces a texture-preserving diffusion model specifically tailored for CBCT-to-CT synthesis. Unlike conventional models, our approach integrates an adaptive wavelet-based high-frequency optimization module to enhance high-frequency details in the CBCT and CT branches, significantly improving feature extraction. Additionally, a dual-mode feature fusion module using an attention-based architecture facilitates effective fusion of CBCT and CT features, allowing CT generation from salient regions while mitigating issues like overexposure and blurring throughout the image generation process. In addition, an innovative edge-aware boundary constraint based on gradient information has been introduced to preserve fine details and sharp transitions in the generated sCT images. The contributions of this research include:

- Introduction of a novel texture-preserving diffusion model tailored for CBCT-to-CT synthesis that enhances sCT image quality by addressing the limitations of the CBCT device.
- Integration of an adaptive high-frequency optimization module prioritizing high-frequency details in CBCT and CT branches, augmenting the quality and interpretability of generated sCT images.
- Development of a transformer-based fusion module for cross-modality features, employing an attention architecture to initiate sCT generation from salient regions and alleviate potential synthesis issues.
- Demonstration of superior performance in multi-site validation compared to existing methods, illustrating enhanced generalization and robustness.

Subsequent sections of this paper delve into our proposed framework in Section 2, detailed experiments and results in Section 3, and comprehensive discussions and summarizations of key methods in Section 4.

2. Methods

2.1. Overview

For DDPM, a forward process ($X_0 \xrightarrow{\text{Diffusion Process}} X_T$) is commonly used in research. Here, X_0 represents the original CT image, X_T depicts the image with progressively superimposed Gaussian noise, and X_t signifies the CT image at step t. This diffusion process can be mathematically expressed as:

$$X_{t+1} = \sqrt{1-\beta_{t+1}}X_t + \sqrt{\beta_{t+1}}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{1}$$

where β_{t+1} determines the noise level in the subsequent step, and ϵ is a variable drawn from a standard normal distribution. The differentiation among various DDPM-based methods primarily lies in the reverse process, which integrates domain knowledge to enhance synthetic images. CBCT images, with low soft tissue contrast, pose a considerable challenge in preserving texture when synthesizing CT images from CBCT.

The method shown in Fig. 1 employs a dual-branch attention-based neural network to denoise noisy CT images. Comprising the Adaptive High-Frequency Optimization (AHFO) module, Dual-Mode Feature Fusion (DMFF) module, Time Embedded (TEB) module, and an edge-aware boundary constraint (EABC) loss function, this framework enhances the traditional diffusion paradigm and augments the model's feature extraction capabilities. It ensures denoising while maintaining consistency with the original CT image. In the reverse diffusion steps, the conditional branch integrates CBCT images to guide the restoration process for the current noisy CT image in the primary CT generation branch. Additionally, a learnable high-frequency optimization module optimizes high-frequency components in both synthetic CT and CBCT features, using the fast Fourier transform (FFT) in the CBCT branch and wavelet transform (WT) in the sCT branch. Subsequently, the DMFF module, employing self-attention, effectively fuses features from both branches. During inference, the network combines random noise, sampled from a standard normal distribution, with input from the CBCT image. This process results in the generation of high-quality sCT images, effectively translating from CBCT to sCT. In general, this approach leverages attention mechanisms, feature optimization, and fusion techniques to denoise CBCT images and produce high-quality sCT images.

2.2. Adaptive high-frequency optimization

To capture high-frequency details in CT and CBCT images accurately, we optimized extracted features in the frequency domain. Considering the characteristics of WT and FFT (Liu et al., 2023): the WT is particularly effective at capturing fine-grained and local frequency structures, more so than the FFT, we applied FFT in the CBCT branch and WT in the CT generation branch, since CT images have more abundant texture features than CBCT images. Then, we decomposes the input feature image x via a transformation function $\mathcal{F}$ (WT or FFT) to derive sub-band feature images, X, as follows.

$$X = \mathcal{F}(x) \tag{2}$$

To optimize the extracted sub-band feature images, reconstruct the CT image's high-frequency details, and emphasize pivotal features, we introduced a learnable weighting matrix, $\mathbf{A}$, as follows:

$$X' = X \odot \mathbf{A} \tag{3}$$

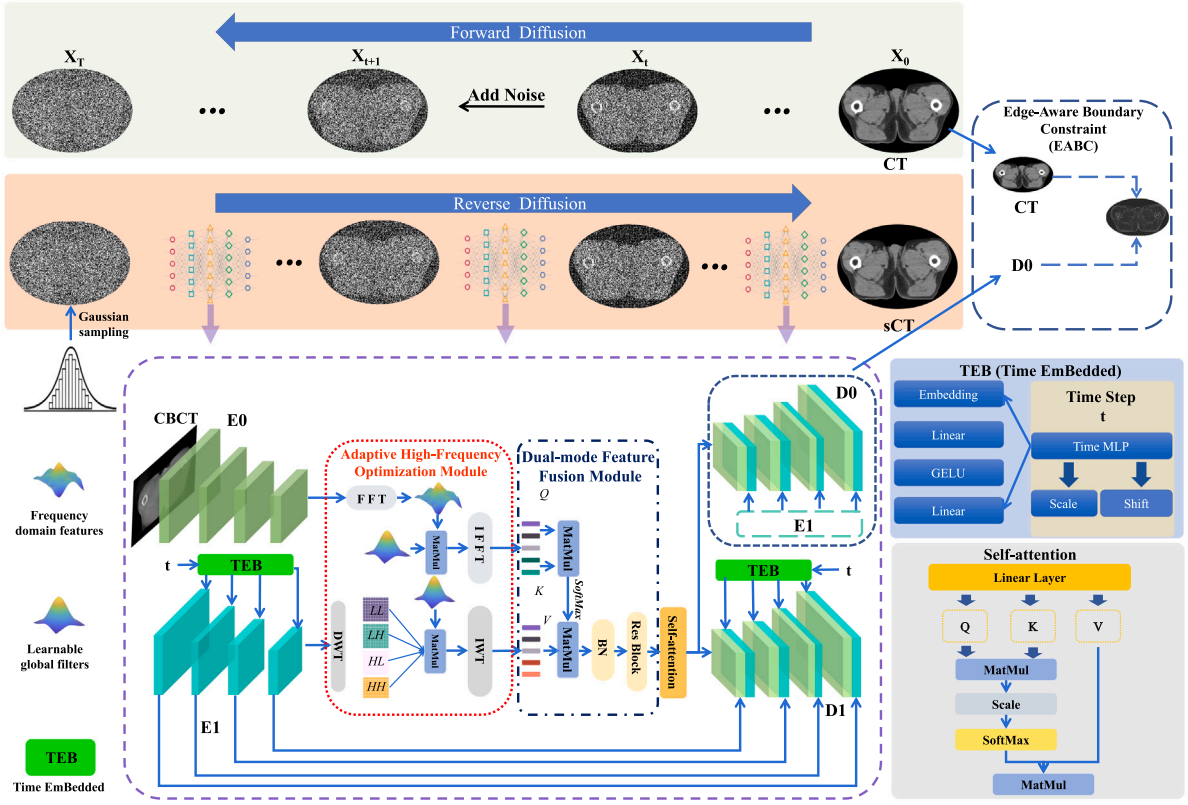

Fig. 1. The overall framework includes entire forward and reverse diffusion process, respectively. **E1** and **D1** are the encoder and decoder in the main branch of sCT generation, respectively. **E0** is the encoder to introduce additional CBCT information. **D0** is the auxiliary decoder for the introduction of the extra loss function of edge-aware boundary constraint.

Here, X' denotes the optimized sub-band feature images, and $\odot$ signifies elemental multiplication. Subsequently, using the reverse transformation function $\hat{F}$, such as IWT or IFFT, we obtained the final output, X_{out}, from the optimized sub-band feature images according to the following equation:

$$X_{\text{out}} = \hat{F}(X') \tag{4}$$

2.3. Dual-mode feature fusion

To more effectively fuse the features of the CT and CBCT branches, we proposed an attention-based feature fusion approach. Specifically, we fused the feature maps between the features of the CT branch **hr**, reconstructed by adaptive WT, and the conditional features in the CBCT branch **fr**, obtained after the transformation of FFT. In this design, **fr** served as the source to generate the query (Q) and the key (K), while **hr** was used to produce the value (V) representation. Through the rearrange method, we could refactor these representations, endowing them with a multi-head attention. To derive the attention weights, we initially computed the similarity between Q and K as Eq. (5), where d denotes the dimension of Q and K. Subsequently, we applied the SoftMax function to this similarity matrix, obtaining the normalized attention weights:

$$\text{Sim}(Q, K) = \frac{Q\,K^T}{\sqrt{d}} \tag{5}$$

These weights emphasize the regions on the CT feature map associated with each position on the CBCT feature map. Then, we utilized these attention weights to optimize V, thus aggregating information from the CT and CBCT feature maps based on the attention scores derived from CBCT:

$$\text{Out} = \text{SoftMax}\left(\text{Sim}(Q, K)\right) V \tag{6}$$

The resulting aggregated output was subjected to batch normalization (BN) and a residual block (ResBlock), which yielded the final fused feature map. Subsequently, a self-attention layer was applied to further emphasize key areas post-fusion.

2.4. Time EmBedded module

In the forward diffusion process, based on the Markov chain, we simulated the data diffusion process by adding different levels of noise at each step. To accurately capture the noise characteristics in the reverse diffusion process at different time steps, we introduced a crucial component, the Time Embedding Module, in the **E1** encoder and **D1** decoder of our diffusion model. Specifically, we inputted signals from different time steps into a Multi-Layer Perceptron (MLP) layer, which is composed of four parts: an Embedding layer, two Linear (fully connected) layers, and a GELU activation layer. Initially, time information was processed through the embedding layer using sinusoidal positional encoding, with the following formula:

$$TE(t, i) = \kappa \left(\frac{\sin\left(\frac{t}{\lambda^{2i/d_{\text{model}}}}\right) + \cos\left(\frac{t}{\lambda^{2i/d_{\text{model}}}}\right)}{2} \right), \tag{7}$$

where t represents the time step. i represents the dimension index. $TE(t, i)$ denotes the encoding value at time step t for the ith index. κ serves as an amplitude value. In this formula, λ is a hyperparameter used to adjust the frequency of time encoding, affecting the periods of the sine and cosine functions, thereby influencing the characteristics of time encoding. d_{model} is the dimension size of time encoding output. After time encoding, two fully connected layers and a GELU activation layer were used for enhanced time information t'. Next t' was divided along the feature dimension into two parts: scale and shift. These

components can dynamically adjust the model's output features, significantly enhancing its ability to handle inherent dynamic noise across various time steps, and finally leading to outstanding performance in the CBCT-to-CT synthesis.

2.5. Loss function

2.5.1. Edge-aware boundary constraint

In order to better capture internal details within the body contour and reduce the interference of background information except for the CT details, as shown in Fig. 1, we employed an auxiliary decoder, **D0**, as the second output, to diligently predict a clean CT image. A gradient-based Laplacian boundary extraction method was used between the output of **D0** and the original CT image to achieve a gradient boundary constraint loss function.

Here, we used the Laplace operator to obtain the second-order gradient, $\nabla^2 I_{D0}(x,y)$, at each pixel point (x,y), of the **D0** decoder output I_{D0}.

$$\nabla^2 I_{D0}(x,y) = I_{D0}(x,y) * L \quad \text{subject to } L = \begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix} \tag{8}$$

Furthermore, binarization was performed to enhance the representation of the edges:

$$B(x,y) = \begin{cases} 1, & \text{if } \nabla^2 I_{D0}(x,y) \geq \theta \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

where θ is a pre-defined threshold to adjust the edge richness. Here, we empirically set $\theta = 0.1$. In this way, we obtained a clear edge map $B(x,y)$, where "1" represents the edge, and "0" represents the non-edge pixel. Similarly, $P(x,y)$, the second-order gradient information of the original CT image, was calculated to be the ground truth for the EABC loss function, $\mathcal{L}_{\text{EABC}}$.

$$\mathcal{L}_{\text{EABC}} = -\sum_{x,y} \Big[B(x,y)\log(P(x,y)) + (1 - B(x,y))\log(1 - P(x,y)) \Big] \tag{10}$$

In this loss function, $B(x,y)\log(P(x,y))$ measures the prediction error when the actual value is 1 (edge), while $(1-B(x,y))\log(1-P(x,y))$ measures the prediction error when the actual value is 0 (non-edge).

2.5.2. Total loss

In this work, apart from EABC, two additional loss functions were also needed to be employed, including $\mathcal{L}_1$ and MSE loss. Since we used the main decoder **D1** of our diffusion model to predict noise and then obtained optimized CT images by subtracting the predicted noise at each step. In each step, the MSE loss is expressed as:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N}\sum_{j=1}^{N}\left(out_{D1}^{j} - \epsilon_{gt}^{j}\right)^2, \tag{11}$$

Here, N represents the number of samples in a batch, out_{D1}^{j} denotes the **D1** output of the jth sample, and ϵ_{gt}^{j} is the noise added to the jth sample at the corresponding step in the forward diffusion process. Also, to address potential outliers in the CT images, we combined the following $\mathcal{L}_{L_1}$ loss:

$$\mathcal{L}_{L_1} = \frac{1}{N}\sum_{j=1}^{N}\left|out_{D0}^{j} - X_{\text{CT}}^{j}\right|, \tag{12}$$

Where N is the batch sample size, out_{D0}^{j} is the **D0** output of the jth sample, and X_{CT}^{j} represents the original jth CT image in the batch samples. The total loss function is formulated as:

$$\mathcal{L}_{\text{total}} = \alpha_{\text{EABC}} \cdot \mathcal{L}_{\text{EABC}} + \alpha_{\text{MSE}} \cdot \mathcal{L}_{\text{MSE}} + \alpha_{L_1} \cdot \mathcal{L}_{L_1}, \tag{13}$$

Here, α denotes the coefficient for each respective loss, signifying the relative importance of that loss in the overall loss calculation. In this study, we set $\alpha_{\text{EABC}} = 10^{-3}$, $\alpha_{\text{MSE}} = \alpha_{L_1} = 1$, and the corresponding abaltion study on how to select these weighting parameters can be seen in Section 3.8.6.

2.6. Accelerated inference

In traditional diffusion models, the inference speed is restricted due to the extensive number of time steps used in both the forward and reverse processes. In our research, to improve the inference speed, we employed a strategy, derived from the Denoising Diffusion Implicit Models (DDIM) (Song et al., 2020), to select a subset of time steps instead of using all time steps as in traditional methods. Thus, this strategy allows skipping certain time steps during the reverse diffusion process. In our work, the subset is determined by evenly selecting specific time steps throughout the entire time sequence, ensuring coverage of key stages of the diffusion process while accelerating the inference. The selection strategy aims to find a balance point, minimizing the number of steps required for inference while ensuring the quality of image generation is not compromised. More specifically, we initiated the process with random Gaussian noise. The CBCT image served as conditional information, facilitating rapid inference to generate the corresponding CT image. At each time step, the model, along with the provided CBCT data, computed the denoised CT image. The iterative strategy for each step is represented by the following equation.

$$\frac{X_{t-\Delta t}}{\sqrt{\beta_{t-\Delta t}}} = \frac{X_t}{\sqrt{\beta_t}} + \sqrt{\frac{1-\beta_{t-\Delta t}}{\beta_{t-\Delta t}}} - \sqrt{\frac{1-\beta_t}{\beta_t}}\epsilon^t X_t \tag{14}$$

In the implementation of our DDIM, we define a predetermined subset $\{\tau_1, \ldots, \tau_S\}$ of time steps. Within this subset, for each pair of consecutive time steps (τ_i, τ_{i+1}), both t and Δt are dynamically determined. Here, t denotes the current time step, and Δt represents the interval to the next selected time step. We initiate the process by establishing a reverse time sequence that extends from the final to the initial time step, and then proceed to sample the sequence by iterating over pairs of adjacent time steps, beginning from the end of this sequence. Consequently, the difference Δt between each consecutive pair of time steps is inherently defined by our strategic selection of time steps. This methodological approach ensures that the time dynamics are appropriately tailored to the specific structural properties of the diffusion process. With this accelerated technique, we achieved rapid inference in the CBCT-to-CT synthesis using the diffusion model while preserving the texture richness of the sCT images.

3. Experimental results

3.1. Dataset

In this work, to train and evaluate the proposed framework, we employed a Pelvis dataset as the training dataset as well as the internal testing dataset. To verify the generalization of the proposed model, multi-site validation on two other external datasets were conducted, including an external Pelvis dataset and an external Head dataset.

3.1.1. Internal Pelvis dataset

79 patients with available CBCT and CT data of the Pelvis were included in our study to train and evaluate the proposed method from the SynthRAD2023 challenge[1] (Thummerer et al., 2023). Subsequently, we randomly divided the 79 patients into training (80%), validation (10%), and testing datasets (10%). In total, the training dataset consists of 5056 slices, the validation dataset has 561 slices, and the testing dataset contains 624 slices.

[1] https://synthrad2023.grand-challenge.org/

3.1.2. External Pelvis dataset

To validate the reliability and generalizability of our proposed method, in addition to the aforementioned internal testing dataset, we incorporated other publicly available Pelvis dataset: the "Pelvic Reference Data"[2] (Clark et al., 2013). This dataset comprises both CBCT/CT images and is only used as an external testing dataset for model evaluation. Before utilizing this CBCT/CT dataset for supervised learning, we performed image registration, as described in Section 3.2, to ensure image alignment.

3.1.3. Different-site data of brain

To assess the efficiency of our proposed method on different tissues, we employed a Head dataset, sourced from the publicly available SynthRAD2023 challenge dataset. Subsequently, the well-trained model, previously trained in the Pelvis training dataset, was used to evaluate the performance in this dataset. The achievement of favorable results on these independent datasets can further verify the generalizability of our proposed method.

3.2. Data registration

In our investigation, given the inherent spatial discrepancies and inconsistent positioning between these two imaging modalities, precise image alignment is of the utmost importance to facilitate accurate comparative analysis. To address this challenge, we employ a state-of-the-art non-linear registration technique known as symmetric normalization (SyN) (Avants et al., 2008). SyN has gained recognition for its effectiveness in reconciling complex anatomical variations, which makes SyN an ideal choice for aligning our CBCT and CT datasets with a shared spatial domain. Specifically, we performed registration by mapping the CBCT images, treated as moving images, onto the CT images, serving as fixed counterparts.

3.3. Evaluation methods

In this work, to comprehensively evaluate the performance of our method, we employed several advanced medical image-to-image translation techniques as comparison methods.

- U-Net (Chen et al., 2020) the commonly used architecture which employs downsampling-upsampling architecture with skip connections to maintain image quality and focuses on contextual information and detail restoration.
- Pix2Pix (Jiang et al., 2022) integrates a U-Net as generator with a PatchGAN discriminator for precise image transformation.
- CycleGAN (Liang et al., 2019a) utilizes two pairs of generators and discriminators, ensuring image consistency and reversibility.
- EaGANs (Yu et al., 2019) incorporates an edge loss function using the Sobel operator.
- ResViT (Dalmaz et al., 2022) uses an aggregated residual transformer module to amalgamate global and local image details with convolutional neural networks.
- CDDPM (Peng et al., 2023) leverages a U-net structure with residual and attention mechanisms, progressively transforming Gaussian noise into the desired CT distribution based on CBCT.
- DDMM (Huy and Quan, 2023) shares noise schedules and latent spaces across different branches provide additional spatial context for enhanced generative capabilities.
- SynDiff (Özbey et al., 2023) combines adversarial diffusion models with unsupervised training, using adversarial projection to refine image sampling accuracy within a cycle-consistent framework.
- FGDM (Li et al., 2023a) guides image diffusion and denoising through frequency domain analysis, combining low-pass and high-pass filtering to preserve and restore mid and high-frequency image details.

Here, we strictly followed their respective literature sources, trained the models exactly according to the methods described in each literature, and employed our training dataset to well-train these models, ultimately obtaining all the results presented in the paper. For the model evaluation, four metrics were employed: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Structural Similarity Index (SSIM), and Peak Signal-to-Noise Ratio (PSNR). Lower RMSE and MAE means more accurate synthesis of CT from CBCT, while higher SSIM and PSNR indicate better sCT image quality.

3.4. Implementation details

In this paper, all experiments involved normalizing the images using their mean and standard deviation, and then restoring the Hounsfield Units using the same mean and standard deviation. Furthermore, to mitigate the computational demands associated with model training, we uniformly resized all images to a uniform resolution of 256×256. Moreover, we implemented our framework using the PyTorch library[3] and performed on two NVIDIA GeForce RTX 3090 GPUs. We set the batch size at 32 and used the AdamW optimizer for network training. In the forward diffusion process, we set the total diffusion step at 1000, while inference sampling was conducted over 60 steps. Furthermore, we adopted the Automatic Mixed Precision capability in PyTorch. To be specific, we employed half-precision floating-point numbers for forward propagation and parameter updates, while we employed single-precision floating-point numbers for the calculation of loss function, which can effectively reduce memory consumption and communication overhead. Lastly, our learning rate scheduling was based on cosine annealing, facilitating periodic adjustments to the rate, accelerating convergence, and easing the model overfitting. After network training, we can evaluate the performance of our model based on one sampling process.

3.5. Experimental results on internal Pelvis dataset

Quantitative analysis: From Table 1, we can see that our proposed model consistently shows superior performance across all metrics. Specifically, our model achieves the lowest RMSE of 46.62 ± 11.77 and MAE of 32.45 ± 7.93, coupled with the highest SSIM of 0.95 ± 0.04 and PSNR of 45.02 ± 6.14. Compared to the baseline from original CBCT, our approach significantly enhances image quality, as indicated by the substantial improvements in all the evaluated metrics. Compared to the U-Net model, our approach substantially reduces RMSE and MAE by approximately 35.75 and 26.75 respectively, while elevating SSIM and PSNR by 0.07 and 13.72. In addition, Pix2Pix, CycleGAN, EaGANs and ResViT, employing adversarial networks, outperform U-Net but do not reach the performance heights of our model. To be specific, Pix2Pix, when compared to our method, exhibits increased RMSE and MAE by 19.30 and 14.36 respectively, while our model presents enhanced SSIM and PSNR by 0.03 and 9.68. A similar pattern is observed with CycleGAN, exhibiting reduced RMSE and MAE by 15.65 and 12.25, respectively, along with increased SSIM by 0.03 and PSNR by 8.2. The EaGANs method, despite its advancements over Pix2Pix, CycleGAN and ResViT, yields a higher RMSE of 56.79 ± 21.06 and MAE of 40.25 ± 14.42, with SSIM and PSNR readings of 0.94 ± 0.062 and 38.26 ± 8.38 respectively, reflecting a less robust performance in comparison to our model. Regarding DDPM-based methods, including CDDPM, DDMM, SynDiff, and FGDM, their performance relative to GAN-based approaches is uneven, showing no clear superiority. However, it is evident that all fall short of the efficacy demonstrated by our proposed method.

[2] https://wiki.cancerimagingarchive.net/display/Public/Pelvic-Reference-Data

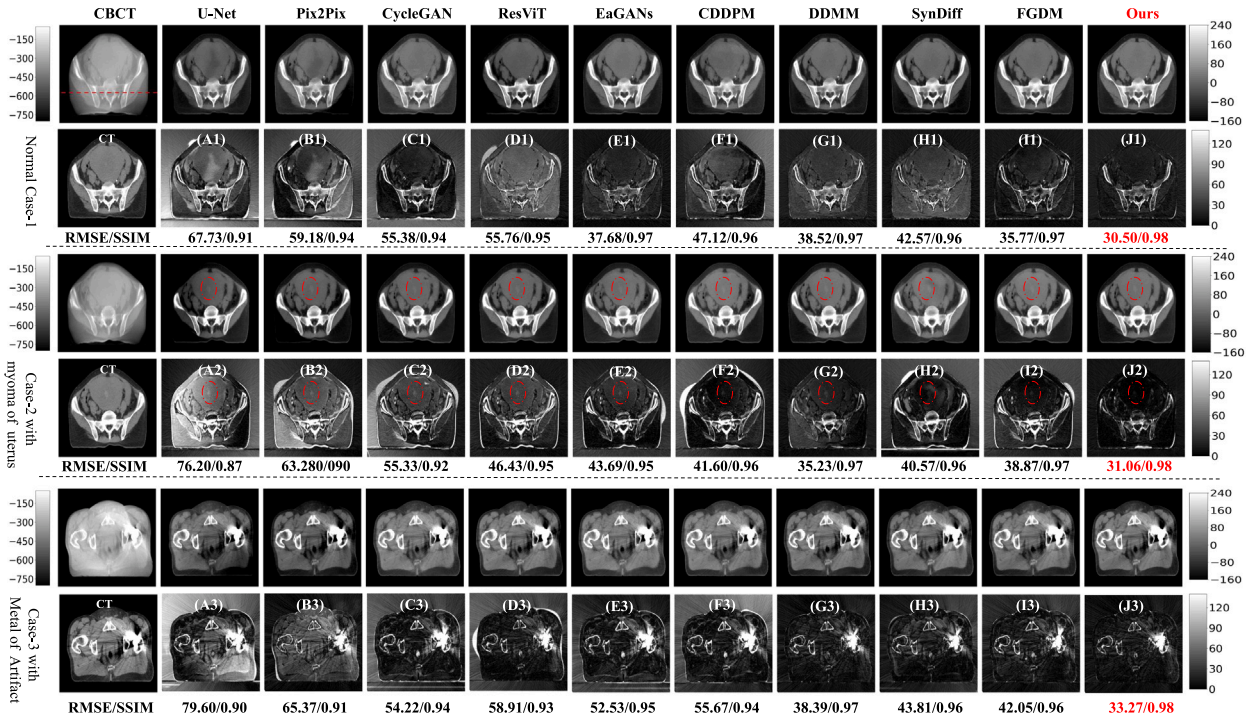

Fig. 2. Visual inspection for experimental results on the internal testing dataset of Pelvis. Here, three representative cases were selected: one is from a healthy individual; one is from an individual with myoma of uterus, marked by the red dashed ellipses; and one from an individual with a hip joint metal implant. (A1-J1), (A2-J2), and (A3-J3) are the absolute difference images between CT and the corresponding predicted sCT. The display window for CBCT is [−800, −50] HU. And the display window for CT and sCT is [−160, 240] HU.

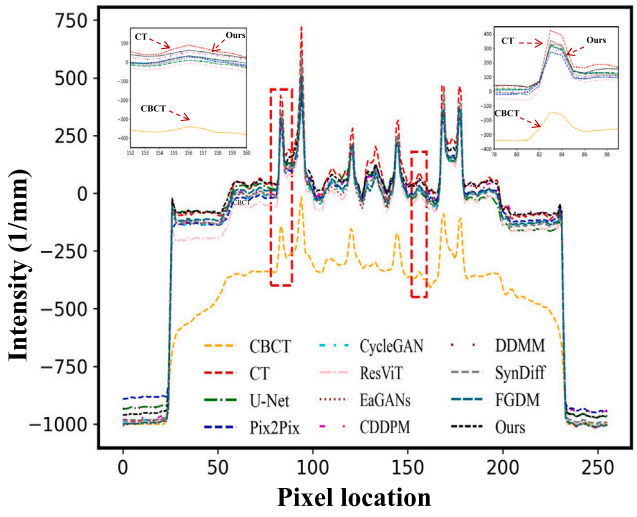

Fig. 3. The 1D intensity profile, depicted in Fig. 2, is obtained by traversing the red-dashed line. The insets within the figure provide a clear visual representation of the superior performance exhibited by the proposed method compared to other methods.

Qualitative analysis: For visual inspection more intuitively, we selected three representative cases for analysis: one case from a healthy individual; one case from an individual with myoma of uterus, which presents pathological changes; and one case from an individual with a hip joint metal implant. Fig. 2 presents the experimental results generated by all the involved methods, along with the absolute difference images (A1-J1), (A2-J2), and (A3-J3) compared to the original CT scans. From these absolute difference images, it is evident that our proposed model outperforms the others in sCT images, and the sCT images generated by our model are more similar to the original CT scans compared to those generated by U-Net, GAN-based, and DDPM-based methods. In the case involving myoma of uterus, minimal residual myoma are visible in Fig. 2 (J2), demonstrating that our method not only facilitates the CBCT-to-CT synthesis but also ensures accurate modality transformation of the lesion features, thereby highlighting the clinical applicability of our approach. In cases with metal implants, although there is variation in the quality of sCT, all methods employed demonstrate robustness against metal artifacts, which significantly degrade the quality of CBCT images, and can produce visually acceptable sCT. Fig. 3 is the 1D intensity profile, marked by the red-dashed line in Fig. 2. We can see that our pixel intensity is closer to the CT intensity shown in the first region of interest (ROI1) and ROI2.

Table 1
Experimental results on the internal pelvis testing dataset.

Model/Metric	Internal testing dataset (SynthRAD2023)			
	RMSE	MAE	SSIM	PSNR
Original CBCT	457.28 ± 57.32	407.82 ± 61.51	0.68 ± 0.03	15.71 ± 4.49
U-Net	82.37 ± 29.90	59.20 ± 18.70	0.88 ± 0.06	31.30 ± 8.57
Pix2Pix	65.92 ± 16.58	46.81 ± 14.92	0.92 ± 0.04	35.34 ± 7.31
CycleGAN	62.27 ± 18.43	44.70 ± 16.30	0.92 ± 0.06	36.82 ± 8.32
ResViT	59.33 ± 25.13	42.81 ± 12.26	0.94 ± 0.07	37.60 ± 9.27
EaGANs	56.79 ± 21.06	40.25 ± 14.42	0.94 ± 0.06	38.26 ± 8.38
CDDPM	60.47 ± 27.08	41.52 ± 13.17	0.94 ± 0.05	37.73 ± 8.86
DDMM	52.08 ± 14.71	36.74 ± 11.06	0.94 ± 0.06	39.08 ± 8.28
SynDiff	61.07 ± 14.01	43.44 ± 16.50	0.93 ± 0.04	35.95 ± 7.88
FGDM	56.08 ± 15.04	39.59 ± 12.44	0.93 ± 0.08	38.86 ± 8.19
Ours	**46.62 ± 11.77**	**32.45 ± 7.93**	**0.95 ± 0.04**	**45.02 ± 6.14**

[3] https://pytorch.org/

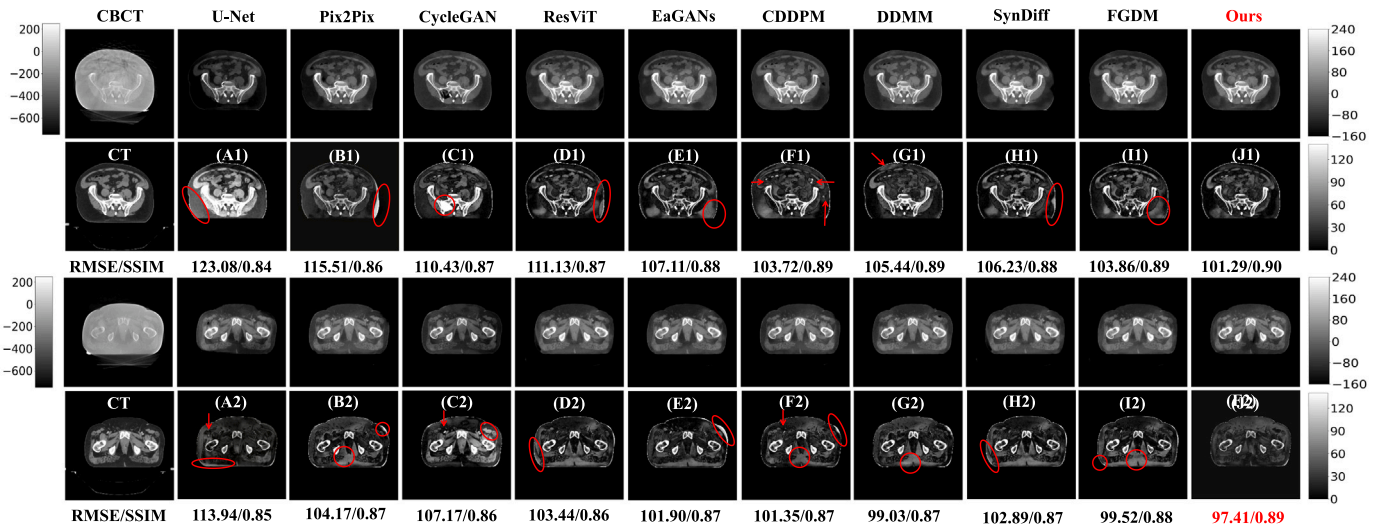

Fig. 4. Visual inspection for experimental results on the external testing dataset of Pelvis. (A1-J1) and (A2-J2) are the absolute difference images between CT and the corresponding predicted sCT. The display window for CBCT is [−750, 250] HU in the external testing dataset. The display window for CT and sCT is [−160, 240] HU.

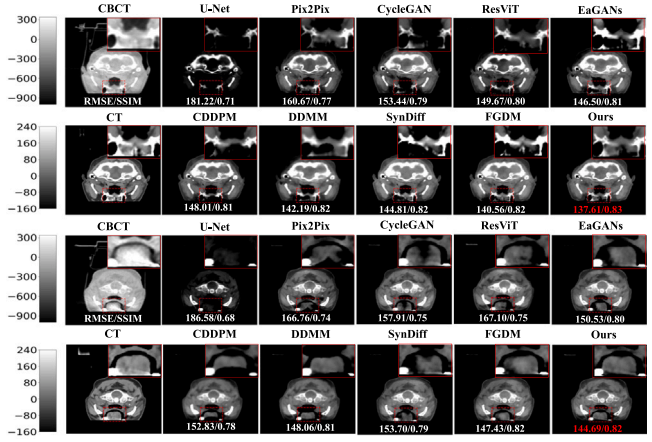

Fig. 5. Visual inspection for experimental results on brain dataset. The second and fourth rows indicate the enlarged region marked by the red-dashed rectangle. The quantitative results was placed in the bottom of the predicted image. The display window for CBCT is [−1000, 340] HU. The display window for CT and sCT is [−160, 240] HU.

3.6. External validation on another Pelvis dataset

External validation is an important method for evaluating the generalizability of a well-trained model. Iantsen et al. (2021) Here, we used an extra external Pelvis dataset, "Pelvic-Reference-Data", for the external cross-validation which only includes Pelvis patient data. After image alignment using SyN registration method, we tested the well-trained model on this external Pelvis dataset. The experimental results can be shown in Table 2. From Table 2, we can see that the performance of all the involved methods from external testing dataset deteriorates than that from internal testing dataset. The potential explanation is that there are some other artifacts in the external testing dataset, such as streak artifacts, *etc.* Anyway, in the external testing dataset, the proposed method still achieves superior performance in terms of all evaluation metrics, showing the strong generalizability of the proposed model. Fig. 4 shows experimental results of two examples. It is evident from the analysis that U-Net, GANs, and DDPMs introduce varying degrees of alterations (lack of details marked by the red arrow and large area of change in gray value marked by the red ellipses) in the internal structure of CT images.

Table 2
Experimental results on the external pelvis testing datasets.

Model/Metric	External testing dataset (Pelvic-reference-data)			
	RMSE	MAE	SSIM	PSNR
Original CBCT	431.29 ± 71.72	374.94 ± 55.09	0.69 ± 0.06	17.87 ± 5.27
U-Net	126.22 ± 27.80	98.36 ± 24.73	0.78 ± 0.08	18.88 ± 5.30
Pix2Pix	116.89 ± 26.65	90.73 ± 18.65	0.81 ± 0.08	19.68 ± 5.93
CycleGAN	114.72 ± 19.00	78.19 ± 16.15	0.82 ± 0.05	21.15 ± 4.12
ResViT	112.09 ± 21.54	73.83 ± 18.60	0.83 ± 0.07	22.33 ± 6.11
EaGANs	110.36 ± 19.67	68.77 ± 18.22	0.84 ± 0.06	22.51 ± 5.39
CDDPM	111.40 ± 21.98	71.51 ± 23.98	0.83 ± 0.06	22.09 ± 6.07
DDMM	109.33 ± 20.06	65.40 ± 22.85	0.85 ± 0.06	23.41 ± 5.78
SynDiff	112.81 ± 22.45	74.32 ± 23.62	0.83 ± 0.07	22.74 ± 6.14
FGDM	108.45 ± 21.13	62.64 ± 22.16	0.85 ± 0.04	23.59 ± 5.52
Ours	**106.05 ± 21.72**	**57.79 ± 23.75**	**0.87 ± 0.05**	**23.87 ± 5.06**

Table 3
Experimental results on the brain dataset.

Model/Metric	Brain dataset (SynthRAD2023)			
	RMSE	MAE	SSIM	PSNR
Original CBCT	493.54 ± 82.39	428.86 ± 63.02	0.66 ± 0.16	10.64 ± 6.47
U-Net	197.65 ± 42.29	139.90 ± 42.36	0.72 ± 0.15	11.89 ± 4.89
Pix2Pix	170.73 ± 37.69	116.98 ± 39.86	0.73 ± 0.13	12.71 ± 3.91
CycleGAN	171.94 ± 41.92	118.63 ± 41.95	0.73 ± 0.14	12.58 ± 4.85
ResViT	165.64 ± 34.21	111.31 ± 35.48	0.73 ± 0.10	14.47 ± 4.48
EaGANs	162.44 ± 43.97	111.21 ± 41.22	0.74 ± 0.12	14.80 ± 5.33
CDDPM	163.45 ± 38.56	109.37 ± 39.50	0.73 ± 0.12	14.80 ± 5.33
DDMM	150.09 ± 40.95	101.29 ± 41.75	0.75 ± 0.12	15.82 ± 5.09
SynDiff	162.52 ± 44.88	109.32 ± 34.31	0.73 ± 0.11	14.79 ± 5.13
FGDM	153.22 ± 39.71	107.49 ± 48.77	0.76 ± 0.10	15.62 ± 4.57
Ours	**146.31 ± 32.80**	**97.88 ± 32.86**	**0.77 ± 0.10**	**16.80 ± 4.38**

3.7. Generalization to different-site data of head

In this study, we specifically investigated the generalization ability of the proposed method across different tissues. To accomplish this, we tested the well-trained framework on a different-site data, which exclusively consists of Head data, while all the related models were trained solely on the internal Pelvis dataset. The quantitative results, presented in Table 3, illustrate the performance of various methods. Notably, our method surpasses other approaches in terms of all metrics. Similar observations are evident from Fig. 5, where the proposed method demonstrates the ability to restore a greater amount of textures within the ROI. In particular, U-Net almost fails on this dataset. Compared to the simple U-Net architecture, in the first sample, our method achieves a notable 24.06% decrease in RMSE and a 16.90% increase in SSIM. Furthermore, we observe a decrease in the second sample of 22.45% in RMSE and a substantial increase of 20.59% in SSIM.

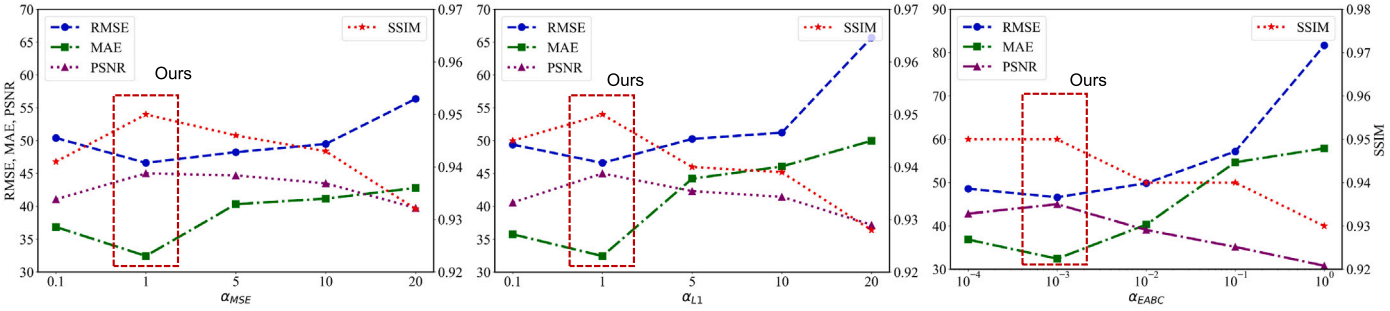

Fig. 6. Evaluation results on the validation dataset during the network training.

Table 4
Ablation study on the proposed generator. Vanilla model means the original model and modified model means using our generator to replace their own generators.

	RMSE	MAE	SSIM	PSNR
Vanilla U-Net	82.37 ± 29.90	59.20 ± 18.70	0.88 ± 0.06	31.30 ± 8.57
Our generator	**76.55 ± 26.89**	**52.55 ± 19.84**	**0.92 ± 0.06**	**35.62 ± 8.19**
Vanilla Pix2Pix	65.92 ± 16.58	46.81 ± 14.92	0.92 ± 0.04	35.34 ± 7.31
Modified Pix2Pix	**60.77 ± 17.40**	**43.90 ± 15.12**	**0.92 ± 0.04**	**36.83 ± 7.81**
Vanilla CycleGAN	62.27 ± 18.43	44.70 ± 16.30	0.92 ± 0.06	36.82 ± 8.32
Modified CycleGAN	**61.74 ± 18.22**	**42.83 ± 16.49**	**0.92 ± 0.04**	**37.40 ± 8.18**
Vanilla EaGANs	56.79 ± 21.06	40.25 ± 14.42	0.94 ± 0.06	38.26 ± 8.38
Modified EaGANs	**54.39 ± 19.57**	**39.77 ± 15.33**	**0.94 ± 0.05**	**38.95 ± 8.41**

3.8. Ablation study

3.8.1. Effectiveness of the proposed generator

To assess the effectiveness of the proposed generator for one-step CBCT-to-CT synthesis, we eliminated the diffusion framework and directly employed our generator to predict sCT images from CBCT scans like what U-Net has done. As demonstrated in Table 4, the generator developed in this study consistently outperforms the standard U-Net, yet falls short of the enhanced outcomes achieved with our combined use of the diffusion framework and the proposed generator, seen in Fig. 2. In addition, we also removed the original generators of GAN-based methods, such as Pix2Pix, CycleGAN and EaGANs, and replaced the original generators in GAN-based frameworks with the proposed generator, unifying the generators used across these methods, making all the involved generators the same. From Table 4, we can see that using our well-designed generator to replace all the vanilla ones will consistently improve the model performance.

3.8.2. Effectiveness of the adaptive high-frequency optimization module

Here, we carried out four extra experiments in Table 5

- M1: the baseline for removing all relevant modules, such as WT, FFT, DMFF, and EABC.
- M2: the modified model of removing modules (FFT, DMFF, EABC)
- M3: the modified model of removing modules (WT, DMFF, EABC)
- M4: the modified model of removing modules (DMFF, EABC)

Table 5 reveals the performance comparisons with the baseline across various configurations. The incorporation of each branch, whether it is the WT or the FFT, results in improved performance. However, it is important to note that these enhancements remain significantly inferior to the performance achieved by M4 with the combination of WT and FFT.

3.8.3. Effectiveness of the dual-mode feature fusion module

In this work, CBCT and CT are distributed in these two branches affecting the performance of the entire model. Here, to evaluate the effectiveness of the proposed dual-mode feature fusion module, we carried out an ablation study that replaced this module with a simple feature concatenation (M5, seen in Table 5). Compared to M7, our method, we can see that without DMFF, RMSE increases 42.25% and MAE 26.90%, also SSIM decreases 3.16% and PSNR 18.26%.

3.8.4. Effectiveness of edge-aware boundary constraint

To assess the effectiveness of EABC in preserving texture, an additional ablation study was conducted by excluding this module from the proposed method (M7). In Table 5, a comparison with our approach (M7) reveals that removal of EABC (M6) leads to a notable increase in both RMSE and MAE, showing increments of 30.72% and 26.69%, respectively. Furthermore, there is a corresponding decrease in SSIM and PSNR, with reductions of 2.11% and 16.08%, respectively. These empirical findings provide compelling evidence for the importance of the edge-aware loss function implemented by the EABC in maintaining the texture fidelity within sCT images.

3.8.5. Ablation of the choice between WT and FFT

Building on the insights from Liu et al. (2024), we can see that the WT is particularly effective at capturing fine-grained and local frequency structures, more so than the FFT. In our task of CBCT-to-CT synthesis, it is pertinent to note that CT images typically contain a higher density of high-frequency components and finer details compared to CBCT images. Consequently, in our adaptive high-frequency optimization module, we have implemented WT within the CT branch to harness these attributes. Conversely, since CBCT images possess fewer high-frequency textures and given the superior computational efficiency of FFT over WT, we opted for FFT in the CBCT branch. The experimental results, as detailed in Table 6, illustrate the impact of performance when substituting WT and FFT between the two branches, including configurations with "all WT" and "all FFT", as well as scenarios where FFT and WT are interchanged ($WT <-> FFT$). These results provide a comprehensive evaluation of the respective strengths of each transform in our application context.

3.8.6. Selection of weighting parameters of loss function

In designing the overall loss function, our goal is to comprehensively consider the accuracy of edge, noise prediction, and CT image reconstruction to ensure that the model preserves important features of the CT images while minimizing the interference from background noise. In our work, the optimal values for each weighting parameter were determined through a series of ablation studies in which individual parameters were perturbed while the others were kept fixed. In Fig. 6, we can see that our selection of weighting parameters is optimal in the vast majority of cases.

3.8.7. Robustness of proposed method on imperfect CBCT

It is well-known that the robustness of a well-trained neural network is a crucial indicator of its generalizability in practical applications. The preliminary results obtained from the test images with metal artifacts in Fig. 2 have validated the robustness of our method. Furthermore, to assess the model's robustness to noise, we introduced varying levels of noise with mean of zero and different standard deviations to all images in the testing dataset and conducted tests using our well-trained model.

Table 5
Comparisons with various modified models for ablation study.

Models	WT	FFT	DMFF	EABC	RMSE	MAE	SSIM	PSNR
M1	✗	✗	✗	✗	86.32 ± 26.89	61.18 ± 14.42	0.90 ± 0.06	30.79 ± 9.03
M2	✓	✗	✗	✗	81.54 ± 28.68	58.52 ± 17.45	0.91 ± 0.05	30.96 ± 7.73
M3	✗	✓	✗	✗	81.75 ± 29.65	55.41 ± 17.28	0.89 ± 0.05	30.99 ± 8.02
M4	✓	✓	✗	✗	70.56 ± 28.29	45.04 ± 16.80	0.91 ± 0.05	35.12 ± 8.53
M5	✓	✓	✗	✓	66.32 ± 26.89	41.18 ± 14.42	0.92 ± 0.05	36.80 ± 9.03
M6	✓	✓	✓	✗	60.94 ± 19.75	41.11 ± 15.18	0.93 ± 0.03	37.78 ± 7.02
M7(Ours)	✓	✓	✓	✓	**46.62 ± 11.77**	**32.45 ± 7.93**	**0.95 ± 0.04**	**45.02 ± 6.14**

Table 6
Ablation study on the choice between WT and FFT.

	RMSE	MAE	SSIM	PSNR
All WT	57.73 ± 15.63	41.21 ± 10.82	0.94 ± 0.05	37.64 ± 7.02
All FFT	63.48 ± 17.59	48.57 ± 10.70	0.91 ± 0.06	35.43 ± 8.76
WT<->FFT	61.29 ± 17.85	45.02 ± 11.61	0.94 ± 0.06	35.72 ± 6.16
Ours	**46.62 ± 11.77**	**32.45 ± 7.93**	**0.95 ± 0.04**	**45.02 ± 6.14**

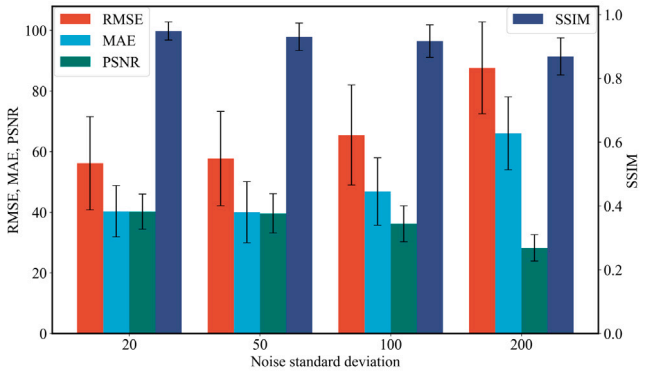

Fig. 7. The impact of noise of varying degrees on the model's robustness.

The results, as depicted in Fig. 7, align with our understanding of neural networks: as the noise level increases, the model's performance correspondingly declines. Specifically, noise within a standard deviation of 100 has a negligible impact on the model, with only minor performance fluctuations. Only when the noise level increases significantly (standard deviation = 200) does the performance experience a relatively substantial decrease.

4. Discussion and conclusion

The synthesis of CT-like images from CBCT data is a growing field, offering a way to take advantage of the benefits of CT while avoiding the restrictions of CBCT technology. This synthesis technique can improve image quality, differentiate between tissues, and provide a more comprehensive understanding of the patient's anatomy, which can lead to more accurate diagnoses and better treatment plans in a variety of clinical settings. Previous studies in this area have mainly focused on GAN-based methods, which use a generator and discriminator to create images. Adversarial training helps the generator create more realistic images, but the difference in capability between the generator and discriminator can sometimes cause pattern collapse, resulting in a lack of image diversity and interpretability. To address these issues, we proposed the use of diffusion models for CBCT-to-CT synthesis. Diffusion models provide a stable and interpretable training approach, allowing for the generation of high-quality CT images while avoiding the problems associated with pattern collapse and interpretability.

In contrast to the end-to-end approaches of U-Net and GANs, our method introduces the utilization of Markov chains to iteratively generate CT images guided by CBCT data from noisy input. This distinctive approach involves employing an auxiliary decoder **D0** within the framework, creating a dual-task architecture unlike other DDPMs. Our modification of the traditional DDPM network incorporates four main components based on U-Net: AHFO, DMFF, TEB, and an additional EABC. To comprehensively evaluate our approach, we conducted a multi-site validation across three independent datasets, including internal and external datasets of Pelvis, as well as a distinct-site dataset of the Head. Through extensive experiments, our method's effectiveness was thoroughly examined across diverse datasets, showcasing its superiority over state-of-the-art models like U-Net, GAN-based, and DDPM-based methods. This comparison emphasizes the generalizability, effectiveness, and versatility of our proposed method.

Within our proposed framework, we integrated FFT and WT to augment feature representation in different branches. Utilizing FFT in the CBCT branch enabled capturing significant global information from a frequency domain perspective, enhancing feature representation. Conversely, in the CT image generation branch from noisy inputs, we employed WT known for adeptly processing local features and extracting detailed information, such as textures (Arivazhagan and Ganesan, 2003). Leveraging the complementary advantages of FFT and WT allowed us to capture both global and local information, improving CT image generation from noisy inputs guided by CBCT. Subsequent to the high-frequency optimization module, the Feature Fusion Module was developed to interact the features of both modalities, demonstrating its superiority in Table 5. Putting emphasis on the importance of boundary information in generation tasks, we proposed an edge-aware boundary constraint loss function. This function, reflecting the direction and magnitude of the information changes through gradients, improved the quality of generated sCT images. By imposing an edge-aware boundary constraint, our method ensured a close alignment of the boundaries in the sCT images with those in the ground-truth CT images. The use of advanced edge detection techniques and edge-guided loss functions preserved fine details and sharp transitions in the generated sCT images, improving visual fidelity and clinical interpretability.

In the present study, our model was designed to facilitate the CBCT-to-CT synthesis in a supervised manner. Although our model has demonstrated commendable performance in the task of image conversion, we also acknowledge its potential and limitations concerning unsupervised domain adaptation. Primarily, our model exhibits significant applicability in conditional generative tasks. Using CBCT as conditional information, our approach is adept at generating the corresponding sCT. This proficiency suggests that, with an adequate dataset of paired images, our model is amenable to extension for conversion tasks between other imaging modalities, such as the transition from CT to Magnetic Resonance (MR) imaging. Furthermore, direct learning from paired data endows our model with superior image generation quality, often exceeding unsupervised methods to preserve structural integrity and fine details, as seen in Table 1, Table 2, and Table 3. However, our model is not without limitations. Chief among these is its reliance on paired datasets, which could be a constraining factor when paired data is costly to annotate or inaccessible. Additionally, the computational demands of our model are substantial due to the noise addition and removal steps inherent in diffusion models, leading to longer training and inference times. This may make the model less suitable for clinical applications that require rapid responses. Another point is that the generated sCT volume may not be coherent between

transverse slices due to the stochastic property of the diffusion model, leading to inconsistency in the coronal and sagittal views. Thus, investigating the methods for consistency in multiple views is a significant direction. In summary, while our model performs well within the current supervised learning paradigm, there is substantial room for improvement in the context of unsupervised domain adaptation. Future endeavors could explore mitigating the dependency on paired data through transfer learning or self-supervised learning techniques and optimize the model to reduce computational costs, thereby enhancing the model's viability in practical clinical settings.

In this study, we introduced a novel texture-preserving denoising diffusion probabilistic model tailored specifically for the CBCT-to-CT synthesis. Our research's key contribution lies in integrating a high-frequency optimization module, a dual-mode feature fusion module, and edge-aware boundary constraints for gradient information preservation, ensuring fidelity in the generated CT images. Additionally, by incorporating time within the encoder and decoder modules, our model enhances focus on target regions of interest. Comprehensive qualitative and visual evaluations exhibit the superior performance of our proposed method compared to U-Net, GANs and other DDPMs. Moreover, our approach enhances interpretability during the generation process. To address the limitation of slower inference speed, future work will explore methods to enhance the sampling speed in CBCT-to-CT synthesis tasks. Our ultimate goal is to generate high-precision images more rapidly, facilitating their application in clinical experiments.

CRediT authorship contribution statement

Youjian Zhang: Writing – review & editing, Methodology, Formal analysis. **Li Li:** Writing – review & editing. **Jie Wang:** Writing – review & editing. **Xinquan Yang:** Writing – review & editing. **Haotian Zhou:** Writing – review & editing, Project administration. **Jiahui He:** Writing – review & editing. **Yaoqin Xie:** Writing – review & editing. **Yuming Jiang:** Writing – review & editing. **Wei Sun:** Writing – review & editing. **Xinyuan Zhang:** Writing – review & editing. **Guanqun Zhou:** Writing – review & editing. **Zhicheng Zhang:** Writing – review & editing, Writing – original draft, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was partially supported by the National Key R&D Program of China (2023YFC2706400, 2022YFC2409000, and 2021YFC3 300500), National Natural Science Foundation of China (82202954, U20A20373), Chongqing Science and Technology Innovation Foundation (CYS23693).

References

Arivazhagan, S., Ganesan, L., 2003. Texture classification using wavelet transform. Pattern Recognit. Lett. 24 (9–10), 1513–1521.

Avants, B.B., Epstein, C.L., Grossman, M., Gee, J.C., 2008. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Med. Image Anal. 12 (1), 26–41.

Barney, B.M., Lee, R.J., Handrahan, D., Welsh, K.T., Cook, J.T., Sause, W.T., 2011. Image-guided radiotherapy (IGRT) for prostate cancer comparing kv imaging of fiducial markers with cone beam computed tomography (CBCT). Int. J. Radiat. Oncol. Biol. Phys. 80 (1), 301–305.

Chen, L., Liang, X., Shen, C., Jiang, S., Wang, J., 2020. Synthetic CT generation from CBCT images via deep learning. Med. Phys. 47 (3), 1115–1125.

Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., et al., 2013. The cancer imaging archive (TCIA): maintaining and operating a public information repository. J. Dig. Imag. 26, 1045–1057.

Dalmaz, O., Yurt, M., Çukur, T., 2022. ResViT: residual vision transformers for multimodal medical image synthesis. IEEE Trans. Med. Imaging 41 (10), 2598–2614.

Ho, J., Jain, A., Abbeel, P., 2020. Denoising diffusion probabilistic models. Adv. neural inf. process. syst. 33, 6840–6851.

Huy, P.N., Quan, T.M., 2023. Denoising diffusion medical models. arXiv preprint arXiv:2304.09383.

Iantsen, A., Ferreira, M., Lucia, F., Jaouen, V., Reinhold, C., Bonaffini, P., Alfieri, J., Rovira, R., Masson, I., Robin, P., et al., 2021. Convolutional neural networks for PET functional volume fully automatic segmentation: development and validation in a multi-center setting. Eur. J. Nuclear Med. Molecular Imag. 48, 3444–3456.

Jiang, Y., Zhang, Y., Luo, C., Yang, P., Wang, J., Liang, X., Zhao, W., Li, R., Niu, T., 2022. A generalized image quality improvement strategy of cone-beam CT using multiple spectral CT labels in pix2pix GAN. Phys. Med. Biol. 67 (11), 115003.

Jiang, Y., Zhang, Z., Wang, W., Huang, W., Chen, C., Xi, S., Ahmad, M.U., Ren, Y., Sang, S., Xie, J., et al., 2023. Biology-guided deep learning predicts prognosis and cancer immunotherapy response. Nature Commun. 14 (1), 5135.

Li, Y., Shao, H.-C., Liang, X., Chen, L., Li, R., Jiang, S., Wang, J., Zhang, Y., 2023a. Zero-shot medical image translation via frequency-guided diffusion models. arXiv preprint arXiv:2304.02742.

Li, L., Tang, Y., Zhang, Y., Li, Z., Zhou, G., Zhou, H., Zhang, Z., 2023b. Federated multi-organ dynamic attention segmentation network with small CT dataset. In: International Workshop on Computational Mathematics Modeling in Cancer Analysis. Springer, pp. 42–50.

Liang, X., Chen, L., Nguyen, D., Zhou, Z., Gu, X., Yang, M., Wang, J., Jiang, S., 2019a. Generating synthesized computed tomography (CT) from cone-beam computed tomography (CBCT) using cyclegan for adaptive radiation therapy. Phys. Med. Biol. 64 (12), 125002.

Liang, X., Li, N., Zhang, Z., Yu, S., Qin, W., Li, Y., Chen, S., Zhang, H., Xie, Y., 2019b. Shading correction for volumetric CT using deep convolutional neural network and adaptive filter. Quantit. Imag. Med. Surgery 9 (7), 1242.

Liu, Y., Lei, Y., Wang, T., Fu, Y., Tang, X., Curran, W.J., Liu, T., Patel, P., Yang, X., 2020. CBCT-based synthetic CT generation using deep-attention cyclegan for pancreatic adaptive radiotherapy. Med. Phys. 47 (6), 2472–2483.

Liu, P., Wu, B., Li, N., Dai, T., Lei, F., Bao, J., Jiang, Y., Xia, S.-T., 2023. Wftnet: Exploiting global and local periodicity in long-term time series forecasting. arXiv preprint arXiv:2309.11319.

Liu, P., Wu, B., Li, N., Dai, T., Lei, F., Bao, J., Jiang, Y., Xia, S.-T., 2024. WFTNet: Exploiting global and local periodicity in long-term time series forecasting. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 5960–5964.

Mututantri-Bastiyange, D., Chow, J.C., 2020. Imaging dose of cone-beam computed tomography in nanoparticle-enhanced image-guided radiotherapy: A Monte Carlo phantom study.. AIMS Bioeng. 7 (1).

Özbey, M., Dalmaz, O., Dar, S.U., Bedel, H.A., Özturk, Ş., Güngör, A., Çukur, T., 2023. Unsupervised medical image translation with adversarial diffusion models. IEEE Trans. Med. Imaging.

Parmar, G., Kumar Singh, K., Zhang, R., Li, Y., Lu, J., Zhu, J.-Y., 2023. Zero-shot image-to-image translation. In: ACM SIGGRAPH 2023 Conference Proceedings. pp. 1–11.

Patel, S., Durack, C., Abella, F., Shemesh, H., Roig, M., Lemberg, K., 2015. Cone beam computed tomography in e ndodontics–a review. Int. Endodontic J. 48 (1), 3–15.

Peng, J., Qiu, R.L., Wynne, J.F., Chang, C.-W., Pan, S., Wang, T., Roper, J., Liu, T., Patel, P.R., Yu, D.S., et al., 2023. CBCT-based synthetic CT image generation using conditional denoising diffusion probabilistic model. arXiv preprint arXiv:2303.02649.

Siewerdsen, J.H., 2011. Cone-beam CT with a flat-panel detector: from image science to image-guided surgery. Nucl. Instrum. Methods Phys. Res. A 648, S241–S250.

Song, J., Meng, C., Ermon, S., 2020. Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502.

Su, X., Song, J., Meng, C., Ermon, S., 2022. Dual diffusion implicit bridges for image-to-image translation. arXiv preprint arXiv:2203.08382.

Thummerer, A., van der Bijl, E., Galapon, Jr., A., Verhoeff, J.J., Langendijk, J.A., Both, S., van den Berg, C.N.A., Maspero, M., 2023. Synthrad2023 grand challenge dataset: Generating synthetic CT for radiotherapy. Med. Phys..

Ujiie, H., Effat, A., Yasufuku, K., 2017. Image-guided thoracic surgery in the hybrid operation room. J. Visualized Surg. 3.

Wang, D., Qin, X., Song, F., Cheng, L., 2020. Stabilizing training of generative adversarial nets via langevin stein variational gradient descent. IEEE Trans. Neural Netw. Learn. Syst. 33 (7), 2768–2780.

Xia, B., Zhang, Y., Wang, S., Wang, Y., Wu, X., Tian, Y., Yang, W., Timotfe, R., Van Gool, L., 2023. Diffi2I: Efficient diffusion model for image-to-image translation. arXiv preprint arXiv:2308.13767.

Yu, B., Zhou, L., Wang, L., Shi, Y., Fripp, J., Bourgeat, P., 2019. Ea-GANs: edge-aware generative adversarial networks for cross-modality MR image synthesis. IEEE Trans. Med. Imaging 38 (7), 1750–1762.