DiffuCE: Expert-Level CBCT Image Enhancement using a Novel Conditional Denoising Diffusion Model with Latent Alignment

Fang-Yi Su, Tzu-Hung Chang, Jung-Hsien Chiang
National Cheng Kung University, Taiwan
{fangyi, lzh107u}@iir.csie.ncku.edu.tw jchiang@mail.ncku.edu.tw

Abstract

Cone-Beam Computed Tomography (CBCT) has garnered significant attention due to lower radiation dosage and faster scanning time, which has been widely used in clinical applications for decades. However, its poor image quality is always challenging to clinical experts. To address this problem, we propose our work **DiffuCE***, a* **Diffu***sion model framework for* **C***BCT* **E***nhancement. The main contributions of our work are three-fold: (1) Increased Generalizability: Our training data exclusively comprises pixel space data, eliminating the necessity for additional imaging machine settings. This emphasizes the model's ability to generalize effectively across diverse conditions. (2) Efficient Training: Rather than starting from scratch, our approach fine-tunes from a well-established foundation model. This illustrates the viability of efficient training strategies for medical image restoration tasks, optimizing resource utilization. (3) Competitive Performance: DiffuCE exhibits outstanding performance, excelling in FID and LPIPS with 0.01 and 36.99 ahead of the second place in the private set. In the public dataset, DiffuCE has a competitive performance compared to other SOTAs. Moreover, in expert assessments, DiffuCE achieves the highest score of 7.06 for overall satisfaction, which is 1.38 ahead of the second place, affirming its performance from a clinical standpoint. Codes are available at* `https://github.com/lzh107u/DiffuCE`

1. Introduction

Cone-Beam Computed Tomography (CBCT) is widely employed in Image-Guided Radiotherapy (IGRT) due to its rapid scanning capabilities, providing radiologists with the latest patient information. However, the poor quality of CBCT images often presents challenges in diagnosis. The important details, such as the boundary between soft tissue and organ, might be blurred by noise, making it difficult for radiologists to use as a reference for IGRT planning. Thus, enhancing CBCT image quality becomes crucial in ensuring the accuracy and safety of IGRT procedures.

CT image reconstruction generally involves translating information between the measurement and image domains. Traditional CT image enhancement heavily relies on prior information, such as imaging settings in the measurement domain [20], and noise distributions like the Gaussian or Poisson distribution [24]. These works focus on the noise patterns that fit the assumptions, which lack generalizability since the assumptions do not always align with practical clinical situations. Additionally, techniques involving the measurement domain often necessitate raw scanning data, such as the raw sinogram or the exact imaging settings, which might not always be available in every circumstance.

Recent works have showcased diffusion models' competitive performance, placing them on par with GANs and establishing them as a new state-of-the-art approach. In line with this, certain medical diffusion models have utilized denoising diffusion probabilistic models [9] or stochastic differential equations [22] on CT [26], MRI [4], and PET images [19], highlighting the potential of diffusion models in restoring medical images. However, these methods primarily operate at the pixel level and lack acceleration in inference, resulting in exceedingly long inference times. Additionally, training a diffusion model from scratch is a resource-intensive process, demanding extensive datasets and substantial GPU resources. These challenges heavily affect the feasibility of deploying novel diffusion-based methods in real clinical scenarios.

To address these challenges, we introduce a pioneering **Diffusion** framework for **CBCT** image **Enhancement** (DiffuCE). This framework effectively eliminates artifacts in the latent space while preserving intricate details using multiple conditional constraint modules. Furthermore, we bridge the gap between the CT and CBCT images in the latent space with an alignment module inspired by CLIP [17], which can stabilize and improve the performance of our framework.

To summarize, our contributions are as follows:

- We have developed a novel framework named DiffuCE, specifically designed to remove artifacts from CBCT images without requiring raw data in the measurement domain. This framework can be a general postprocessing module in any existing CBCT image enhancement pipeline.

- Our framework incorporates an alignment module to bridge the gap between CT and CBCT images within the latent space. This innovation allows our framework to operate without the necessity of paired training data, a rarity in CBCT image enhancement methodologies.
- In the private dataset, our framework gets the best score in FID and LPIPS, demonstrating its ability on par with other methods; In the public dataset, our framework demonstrates competitive performance compared to other SOTA methods, while showing its generalizability across different datasets.

2. Related Works

CBCT Image Enhancement. CT image enhancement involves practical clinical concerns such as radiation dosage [3], scanning angles [2] [20], and so on. Each scanning condition will result in different noise patterns. With enough prior information, a noise pattern can be removed along with digital signal processing or a statistical-based algorithm. [24] propose a method based on digital signal processing on the sinogram with statistical information. With the huge success of CNN-based neural networks in natural image tasks, various architectures [7] [15] [25] [14] [3] are proposed, showing competitive ability of CNN networks in medical image enhancement. However, these methods often lack generalizability, fitting only on a designated noise pattern or imaging setting.

Diffusion Models. The diffusion model is a generative model based on diffusion and denoising processes. In the diffusion process, the noise is gradually added to the data sample, building a path that connects the data and prior distribution. In the denoising process, the model learns a proper way to gradually remove the noise from the data sample. [9] proposes a diffusion process based on the Markov Chain, and [22] proposes a method based on Stochastic Differential Equation(SDE), known as score-based diffusion models. [5] controls the generation process with a classifier, and [10] later introduces a way to control the generation without a classifier. To speed up the time-consuming inference of diffusion models, [21] adapts fewer denoising steps, and [28] [16] propose new mathematical solvers to boost up the inference speed without extra training. [18] utilizes an encoder to compress the image into latent space for denoising, which reduces the computation cost of diffusion models for large image size. Along with the rapid success of diffusion models in natural image processing, several attempts [19] [4] [26] at medical image enhancement have been proposed. However, these networks are trained on pixel space instead of latent space, which needs more computation resources during inference. Moreover, these works are trained from scratch instead of fine-tuned from a pre-trained foundation model, making it hard to obtain for those clinical teams without help from computer science experts. To overcome these challenges, we propose a method based on a pre-trained latent diffusion model, which reduces the computation costs during the inference phase and provides a stable fine-tuning pipeline to lower the difficulty of obtaining.

3. Method

The objective of this research is to generate high-quality synthesized CT images based on the given low-quality CBCT images. Assume $l_{i=1,2,...m}$ indicates the i-th low-quality CBCT image sampled from the CBCT image set L. For each CBCT image, there is:

$$l_i \sim L, i = 1, 2, ..., m; l_i \in R^{c\times h\times w}, \tag{1}$$

, and the objective of our research is to train a framework f that:

$$h_i = f(l_i); h_i \in R^{c\times h\times w}, \tag{2}$$

where h_i is the corresponding high-quality version of l_i, a synthesized CT image.

3.1. Preliminary Study: DDPM

To address this problem, we choose the Denoising Diffusion Probabilistic Model(DDPM) [9] as the backbone. The DDPMs connect the data distribution to a simple distribution, like isotropic Gaussian distribution, with a diffusion process, also known as the forward process. Given a high-quality CT image x_j sampled from the CT image set H. For each CT image, there is:

$$x_j \sim H, j = 1, 2, ..., n; x_j \in R^{c\times h\times w}, \tag{3}$$

, and the data distribution H can be connected to the Gaussian distribution with the forward process that:

$$q(x_{j,t}|x_{j,t-1}) := N(x_{j,t}; \sqrt{1-\beta_t}x_{j,t-1}, \beta_t I) \tag{4}$$

where $t \sim [1, T]$ is timestep and $\beta_1, ..., \beta_T, \beta_t \in (0, 1)$ are fixed variances, both belonging to the scheduler of the DDPM. For a sample x_j at each timestep t, there is:

$$q(x_{j,t}|x_{j,0}) := N(x_{j,t}; \sqrt{\bar{\alpha}_t}x_{j,0}, (1-\bar{\alpha}_t)I), \tag{5}$$

$$x_{j,t} = \sqrt{\bar{\alpha}_t}x_{j,0} + \sqrt{1-\bar{\alpha}_t}\epsilon, \epsilon \sim N(0, I) \tag{6}$$

where $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{k=1}^{t} \alpha_k$. The forward process can be viewed as a linear combination of the data

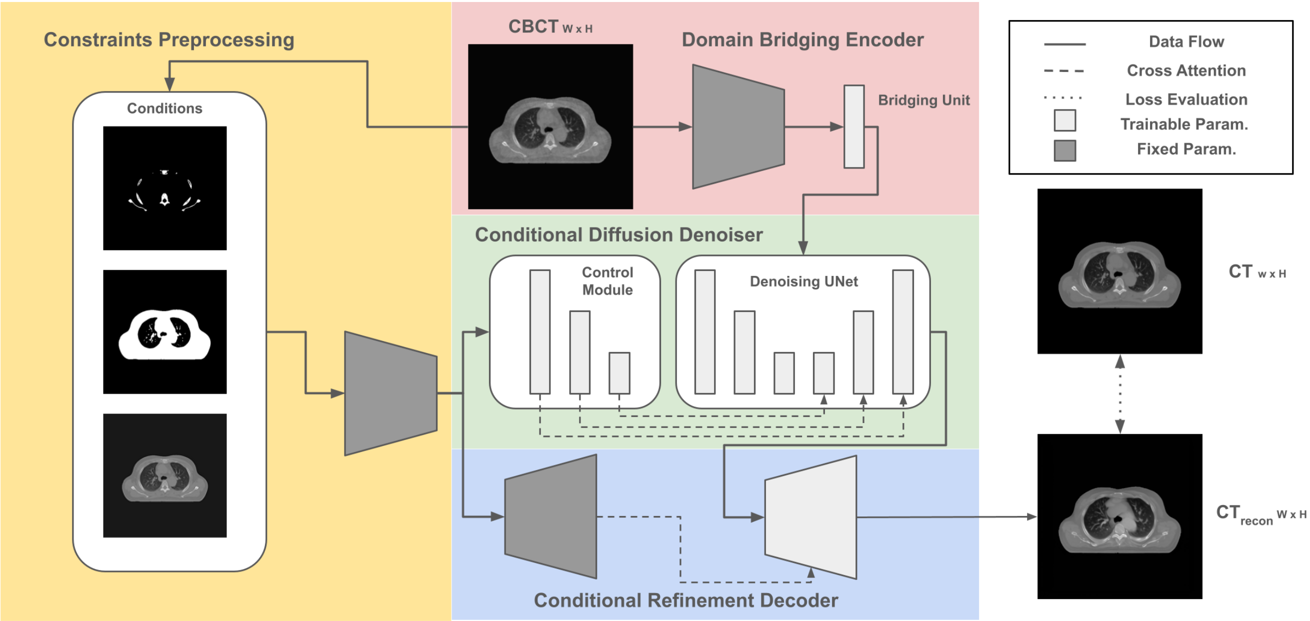

Figure 1. **Overview of the DiffuCE framework.**The DiffuCE framework has four components: 1) Constraints Preprocessing generates conditions from the CBCT input for guidance; 2) Domain Bridging Encoder (DBE) encodes the CBCT image into a noisy-CT-like latent embedding; 3) Conditional Diffusion Denoiser (CDD) removes noise while preserving detail with the help of constraints; 4) Conditional Refinement Decoder (CRD) decodes the clean latent embedding and constraints back to pixel space for reconstruction. LoRA fine-tuning optimizes all parameters. The final output is evaluated against CT ground truth, with loss terms detailed in Section 3.4.

sample x_j and the random noise ϵ. These samples with different ratios of random noise form a path between the Gaussian distribution and the data distribution in the latent space.

It is easy to generate a data sample by sampling a data point from the Gaussian distribution and traversing along the path from the Gaussian distribution to the data distribution, which is called the backward process. For a sample x sampled from the Gaussian distribution $N(0, I)$, there is:

$$p_\theta(x_{t-1}|x_t) := N(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)), \quad (7)$$

where the $\mu_\theta(x_t, t)$ is substituted with a model to predict the noise in the given timestep as follows:

$$\mu_\theta = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha_t}}}\epsilon_\theta(x_{j,t}, t)), \quad (8)$$

where the $\epsilon_\theta(x_t, t)$ is the model that predicts the noise in the given timestep. To learn the backward process, the model $\epsilon_\theta(x_t, t)$ is trained with the objective function as follows:

$$L := E_{t\sim[1,T], x_{j,0}\sim H, \epsilon\sim N(0,I)}[||\epsilon - \epsilon_\theta(x_t, t)||^2], \quad (9)$$

which is an MSE loss that evaluates the prediction of the added Gaussian noise.

Note that the notation of the data sample in the forward process is different from the one in the backward process. In the forward process, the sample x_j is picked from a set of existing CT images, which has its number j; In the backward process, the data point x is sampled from a Gaussian process, which is not an existing data point with specific number.

3.2. Conditional Diffusion Denoiser

Derived from DDPM [9], the Conditional Diffusion Denoiser (CDD) aims to generate high-quality CT images. First, to ensure the quality and stability of diffusion models for processing medical images, we plug LoRA [11] modules into the pre-trained latent diffusion model and fine-tune the entire framework with CT images. Second, to avoid from losing any information, the low-quality CBCT images are directly used in the generation process, leading to an image-to-image generation. Third, inspired by ControlNet [27], multiple guidance modules are also integrated into the CDD to enable more sophisticated constraints during the reverse process. With the help of the LoRA [11] and guidance modules, details from the original input can be preserved as much as possible during the denoising process.

Constraints Preprocessing. To provide solid guidance during the denoising process, clear and deterministic conditions are necessary. Based on clinical knowledge, features such as the lucent area and bone in data can be extracted with specific threshold values in the Hounsfield Unit (HU). The theory beyond HU is the x-ray absorptivity of different

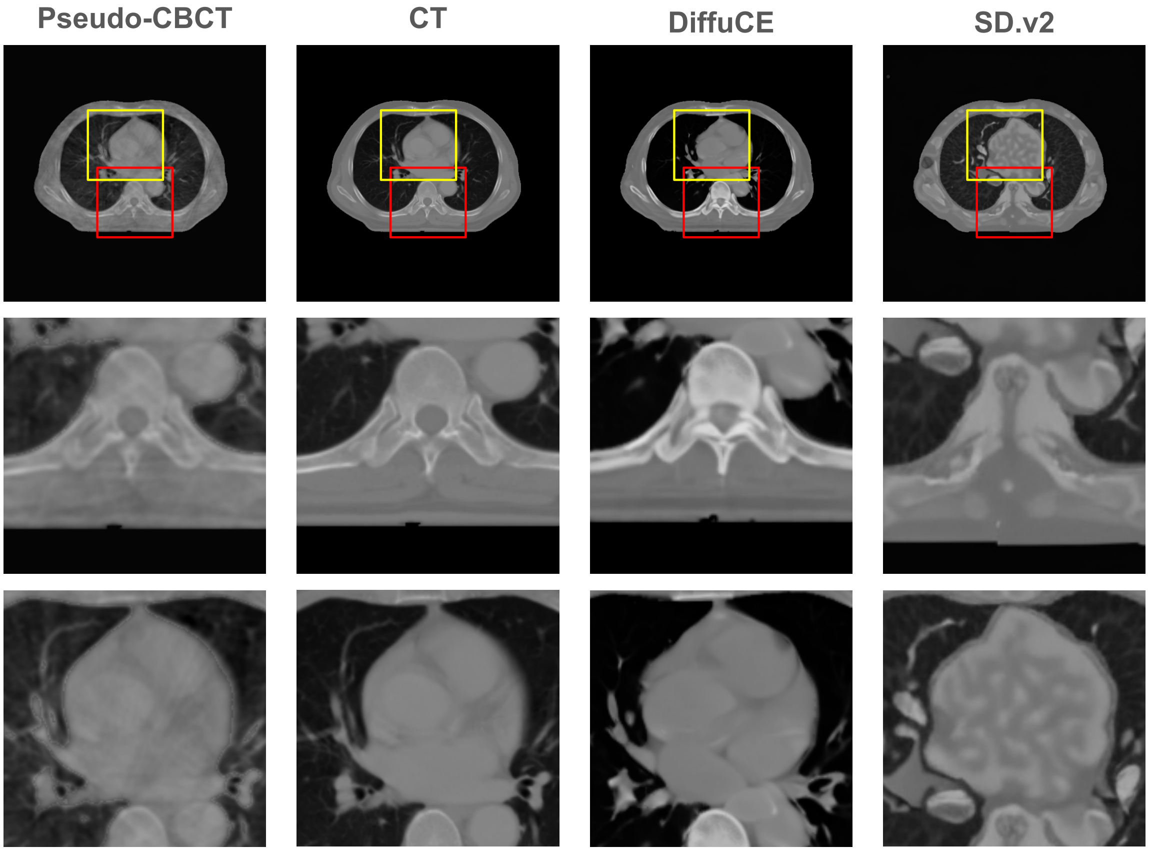

Figure 2. **Case study of DiffuCE and SD.v2 [18].** The impact of condition guidance modules is evident in this figure. The regions outlined by a red line and a yellow line are presented in the second and third rows, respectively. DiffuCE, employing additional conditions as inputs, demonstrates the ability to preserve more structure and texture compared to SD.v2 [18], which solely utilizes pseudo-CBCT as its input.

types of tissue, and it is a general attribute irrelated to the imaging machine, indicating that features extracted by HU values are machine-independent. We use bone, lucent area, and the low-frequency part of input by wavelet transform as a constraint. Unlike the features extracted by HU values that highlight specific parts of an image, the low-frequency feature provides a general view of the given CT image, preventing the CDD from producing incorrect patterns.

Conditional Guidance. Conditional guidance in diffusion models refers to providing additional guidance during denoising iterations to suppress the diversity of the model. The concept of classifier-free conditional guidance in diffusion models is proposed in [10], indicating that the output of diffusion models θ_{diff} can be guided by the given input x and condition c with the combination,

$$\hat{\epsilon_\theta}(x, c) = (1 + w)\epsilon_\theta(x, c) - w\epsilon_\theta(x) \tag{10}$$

, where the $\hat{\epsilon_\theta}$ is the score prediction with conditional guidance with ϵ_θ being a score estimator and w being a scalar for adjusting the combination between conditional and unconditional score estimation. In DiffuCE, we use the constraints preprocessing conditions to guide the CDD.

In short, with conditional guidance, the CDD can generate high-quality synthesized CT images close enough to the input low-quality CBCT images. The complete training algorithm is shown in Algorithm 2 of supplementary materials.

3.3. Domain Bridging Encoder

In Equation 8 and 9, the diffusion model can eventually build a trajectory from the Gaussian distribution to the data distribution. In this research, the trajectory we create in the CDD is between the Gaussian and data distribution. However, the image used in the forward process is the CBCT image instead of the CT image, which makes two distinct trajectories in one generation task, one for CBCT-Gaussian and one for CT-Gaussian. We name this challenge **Distribution Gap** and discuss it in the supplementary materials.

To address the challenge, we propose the Domain Bridging Encoder (DBE) bridges the CBCT and CT image distribution, transforming samples from one to the other. Aside from streak artifacts, the basic layout and contour of a CT image and its corresponding pseudo-CBCT image are the same, indicating that the low-frequency components of these two modalities are alike. Based on this finding, we assume that removing different high-frequency components can partially compensate for the gap between the CT and CBCT image distribution. We validate this assumption in the Experiment Section. In training, we add the Gaussian noise to both CT and pseudo-CBCT images and then train a learnable projector θ_{DBE} that maps noisy CBCT samples to noisy CT distribution.

Alignment Loss. The Alignment Loss is applied to measure the distance between the fixed CT distribution and projected CBCT distribution during training, calculating the distance based on auto-correlation matrices from these two distributions. Given a pair of CT and pseudo-CBCT latent embeddings $\epsilon_{ct}, \epsilon_{cbct} \in \mathbb{R}^{n\times n}$, the auto-correlation matrix can be computed as the following:

$$R_{ct} = \mathrm{E}((\epsilon_{ct} - \mu_{ct})(\epsilon_{ct} - \mu_{ct})^T), \quad (11)$$

$$R_{cbct} = \mathrm{E}((\epsilon_{cbct} - \mu_{cbct})(\epsilon_{cbct} - \mu_{cbct})^T). \quad (12)$$

with $\mu_{ct}, \mu_{cbct} \in \mathbb{R}^{n\times n}$ calculated from all elements in $\epsilon_{ct}, \epsilon_{cbct}$ and extended to the same shape. The Alignment Loss can be further computed by

$$L_{align} = (H(\epsilon_{ct}, target) + H(\epsilon_{cbct}, target))/2, \quad (13)$$

with $target = (R_{ct} + R_{cbct})/2$, and the H is referred as cross entropy.

In brief, the DBE encodes the CBCT image to the latent space and transforms it into the CT distribution to maintain domain consistency between the encoder and latent diffusion models. The complete training algorithm is shown in Algorithm 3 of supplementary materials.

3.4. Conditional Refinement Decoder

In Equation 9, the loss of diffusion model mainly focuses on latent embeddings, which can't directly evaluate the pixel-level accuracy in the image domain. In the realm of clinical, every pixel matters. Any mis-generation will lead to serious consequences. To address this issue, we propose the Conditional Refinement Decoder(CRD) to control the details of the synthesized images within the pixel level.

Conditional Branch. To accurately control the decoding sequence, the information obtained from constraints is integrated into the reconstructed latent embedding by the conditional branch. Although the control modules in CDD perform similar functions, the constraints from these modules are applied in the latent space rather than pixel space. This often results in less accurate details in the outputs when evaluated at the pixel level.

In detail, the CRD integrates the information from conditions to the reconstructed CT image in every network block as follows:

$$\hat{\epsilon_c} = \theta_{branch}(\epsilon_c, \emptyset), \quad (14)$$

$$\epsilon_{\hat{recon}} = \theta_{main}(\epsilon_{recon}, \hat{\epsilon_c}) \quad (15)$$

where θ_{branch} provides embedding $\hat{\epsilon_c}$ from condition ϵ_c and θ_{main} decodes the latent embedding ϵ_{recon} into reconstructed CT image $\epsilon_{\hat{recon}}$ with the help of $\hat{\epsilon_c}$.

Adaptive DualScope Loss. To evaluate the performance of the CRD during training, we propose using the Adaptive DualScope Loss (ADL) to catch the reconstruction quality on both large and small scales. In the following part, the H denotes the high-quality CT image ground truth dataset, and the $\hat{H}$ denotes the reconstructed image dataset. First of all, the region-wise loss L_{reg} evaluates the reconstruction of bone and lucent area. For each condition c_i, there is a mask operator $\mathcal{M}_i$ blocking out the irrelevant area

$$L_{reg} = \mathbb{E}_{x\sim H, \hat{x}\sim \hat{H}} \sum_{i=1}^{k} \|\mathcal{M}_i(x) - \mathcal{M}_i(\hat{x})\|^2, \quad (16)$$

enabling the evaluation can focus on the designated area instead of the whole image. Secondly, the perceptual loss [13] aims for the evaluation of the whole image. In comparison to the traditional loss being sensitive to the pixel value, the perceptual loss [13] evaluates the reconstruction quality by calculating the similarity of embeddings obtained from the CT and reconstructed image with the CNN-based network $\mathcal{N}$

$$L_{percept} = \mathbb{E}_{x\sim H, \hat{x}\sim \hat{H}} \|\mathcal{N}(x) - \mathcal{N}(\hat{x})\|^2, \quad (17)$$

being closer to human visual recognition. Last but not least, we utilize a discriminator $\mathcal{D}$ to distinguish x_{ct} and x_{recon}. With the participation of $\mathcal{D}$, the training procedure of the CRD can be in a GAN-based style

$$L_{adv} = \mathbb{E}_{x\sim H}[log\mathcal{D}(x)] + \mathbb{E}_{\hat{x}\sim \hat{H}}[log(1 - \mathcal{D}(\hat{x}))] \quad (18)$$

adversarially improving the ability of the CRD. In particular, we adaptively scale the $L_{reg}, L_{percept}$ to the value as the largest one at every optimization step, balancing the contribution of each objective.

In summary, the CRD decodes samples from latent space to pixel space with constraints for pixel-level detail preservation. The training algorithm is shown in Algorithm 4 of supplementary materials.

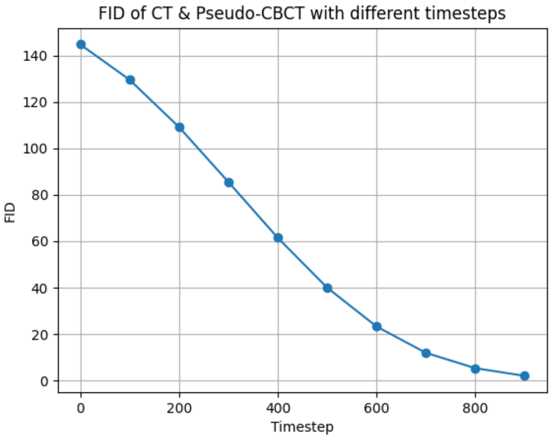

Figure 3. **FID of CT and Pseudo-CBCT image dataset with different timesteps.**

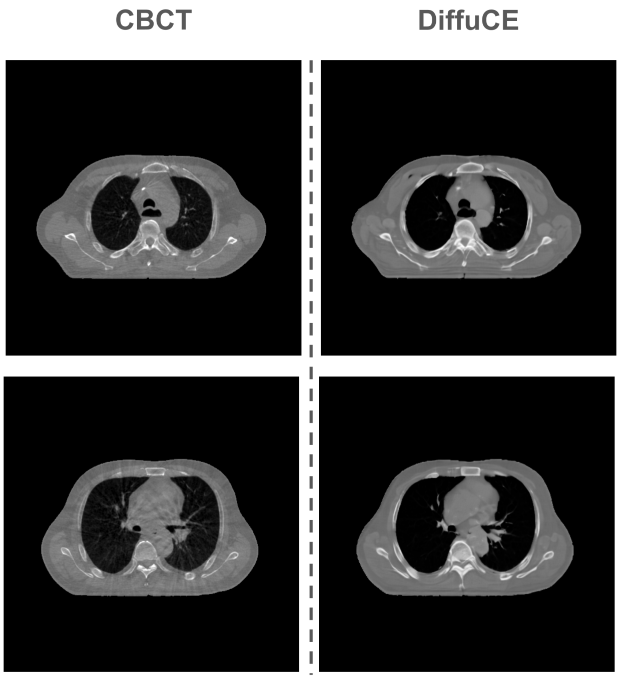

Figure 4. **Case Study: Real CBCT processed by DiffuCE.**

4. Experiments

4.1. Dataset Evaluation

Datasets. We evaluate DiffuCE on two datasets: **SynthRAD2023 CBCT-to-CT Pelvis Dataset** [12] [23](SynthRAD set) and a **private dataset from the collaborated medical center**(Private set). SynthRAD set is a pair-wise dataset mainly focusing on medical image translation, including MRI-to-CT and CBCT-to-CT. We pick the pelvis dataset from the CBCT-to-CT task, containing 180 3D CBCT and CT image pairs collected from three different centers. We split 150 pairs as the training set and 30 pairs as the validation set.

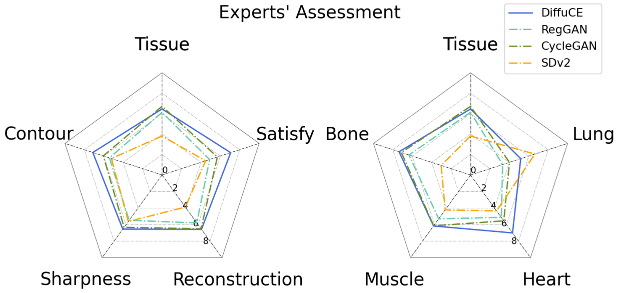

Figure 5. **Experts' Assessment.** The answers from different sets of chest CBCT images and the experts are averaged in this figure.

In the private dataset, the interval between the CT and CBCT scanning is about one month, containing 4177 2D CT images and 2816 2D CBCT Chest images. We use 3000 CT images as the training set for the CDD and 343 CT images for the training of the DBE and CRD. We evaluate our model on the rest of 1177 CT images and randomly pick CBCT images to conduct qualitative research.

Metrics. There are five metrics used in our research: **MAE**, **PSNR**, **SSIM**, **LPIPS** [29], and **FID** [8]. MAE, PSNR and SSIM are used for traditional pixel-level evaluation, while LPIPS [29] is used for human visual recognition similarity. The FID [8] is used to evaluate data sample distribution.

Quantitative Result of the Private Set. The results of the private set are shown in Table 1, revealing that our method, DiffuCE, has a competitive ability in the CBCT enhancement task. Specifically, DiffuCE gets the first place on LPIPS [29] and FID [8], indicating that the reconstructed CT image from DiffuCE not only has a higher quality in human visual perception but also well capture the distribution of training dataset compared to other competitors. However, conventional pixel-level metrics such as PSNR and SSIM only get second place, indicating that pixel-level diversity is still a critical challenge in our research. It's worth noting that the Stable Diffusion V2 (SD.v2) [18] attains the second place in the FID [8], showcasing that diffusion models can effectively match the distribution of the training dataset. However, it is observed that during denoising, the reconstruction samples lack control, leading to highly distorted outcomes. We will discuss this finding in Section 4.3. A case study is shown in Figure 2.

Quantitative Result of the SynthRAD Set. The re-

Model	PSNR↑	SSIM↑	LPIPS↓	FID↓
Baseline	33.43	0.74	0.05	88.89
GAN [6]	15.75 ± 1.57	0.75 ± 0.028	0.02 ± 0.012	275.09
RegGAN	22.32 ± 1.71	0.88 ± 0.024	0.12 ± 0.024	142.11
CycleGAN [30]	27.01 ± 2.04	**0.93 ± 0.018**	0.07 ± 0.015	100.16
CDGAN [1]	**28.97 ± 0.99**	0.88 ± 0.037	0.11 ± 0.01	104.17
SD.v2 [18]	27.49 ± 1.21	0.70 ± 0.25	0.094 ± 0.021	92.55
Ours	27.78 ± 1.40	0.90 ± 0.022	**0.06 ± 0.015**	**55.56**

Table 1. **Pseudo-CBCT dataset Quantitative Results.** The best performance on each metric has been marked with bold letters, and the second place has been underlined. The **baseline is the direct comparison between CT and pseudo-CBCT raw images**, and note that the baseline has been excluded from the comparison.

Model	MAE(HU,↓)	PSNR(dB,↑)	SSIM(↑)
GEneRaTion	55.50 ± 11.00	30.48 ± 1.72	0.897 ± 0.033
KoalAI	56.13 ± 12.06	30.11 ± 1.89	0.897 ± 0.034
SMU-MedVision	49.95 ± 11.78	30.79 ± 2.00	0.906 ± 0.036
FGZ Medical Research	60.65 ± 12.56	29.67 ± 1.71	0.879 ± 0.039
Stratified baseline	69.99 ± 18.93	28.65 ± 2.25	0.837 ± 0.057
Water baseline	344.26 ± 125.32	17.97 ± 2.08	0.546 ± 0.149
Ours†	132.73 ± 50.49	23.66 ± 1.47	0.82 ± 0.038
Ours	132.31 ± 50.13	23.67 ± 1.46	0.82 ± 0.038

Table 2. **SynthRAD2023 Task2 Pelvis Dataset Quantitative Results.** Among all the works, the **FGZ Medical Research** is also a diffusion-based approach directly trained on pixel-level data. Further information on works listed in the table can be found in [12]. † indicates our framework with CRD trained on a private dataset.

sults of the SynthRAD2023 CBCT-to-CT pelvis dataset are shown in Table 2, with mean value and one standard deviation. Besides our work, we provide part of competitors in [12] on task2, which are the best works on different backbones. Our work outperforms the Water baseline, and almost reaches the Stratified baseline on SSIM, showing that our work still has a competitive ability compared to the best works on the leaderboard. Interestingly, the ability of our framework with CRD trained on the private set, marked with †, is close to the one trained on the SynthRAD set. It might indicate that the different data distributions, caused by different machines and noise patterns, are bridged to the same distribution, representing that DiffuCE can capture the general features of high-quality medical images across different tasks. We will discuss this finding in supplementary materials.

Distribution Gap The results in Figure 3 support our assumption about the Distribution Gap. In early timesteps, the FID between two types of CT images is large, which means the trajectories created by these two datasets are different. With stronger Gaussian noise in larger timesteps, the FID of two datasets is reduced. With a moderate level of noise, the complexity of distribution mapping is lower enough to be achieved by minimal training parameters like LoRA, and meanwhile, the information hasn't been distorted too much. In the implementation, we choose 300 as the setting of timestep in the DBE.

Qualitative Result. Conventional metrics such as MSE and PSNR, commonly used in computer vision tasks, provide an objective measure of performance primarily focusing on pixel space. However, these metrics have been shown to inadequately capture image quality from a clinical standpoint. Clinical experts prioritize structural or semantic features, such as the shape of organs within task-dependent regions of interest (ROI), often disregarding content outside the ROI. As a result, ordinary metrics based solely on pixel values may fail to accurately reflect the effectiveness of our framework. To vividly illustrate our advancements, we have presented some case studies across various experimental scenarios in Figure 4 and supplementary materials. Moreover, we go ahead and introduce the expert assessment with questionnaires in the next section.

Experts' Assessment. To assess our reconstruction capability from a clinical perspective, we engage five radiologists from the local medical center in an evaluation. The assessment involves chest CBCT images from 10 patients, each CBCT image accompanied by several reconstructed samples generated from various models, including GAN-based and diffusion-based methods. The questionnaire, arranged with the CBCT images and their corresponding reconstructions, aims to evaluate nine different metrics using a Likert scale. Results in Figure 5 indicate that the DiffuCE

Model	PSNR↑	SSIM↑	LPIPS↓	FID↓
w/o DBE	$\mathbf{28.80 \pm 1.43}$	$\mathbf{0.91 \pm 0.023}$	$\mathbf{0.055 \pm 0.013}$	53.33
w/o CDD	27.70 ± 1.43	0.89 ± 0.024	0.069 ± 0.018	78.66
w/o CRD	26.42 ± 0.70	0.44 ± 0.058	0.088 ± 0.020	**52.57**
Ours	27.78 ± 1.40	0.90 ± 0.022	$\mathbf{0.06 \pm 0.015}$	55.56

Table 3. **Ablation study**. All of the settings in the ablation study are evaluated on the pseudo-CBCT and CT image dataset without additional training. The best performance on each metric has been marked with bold letters, and the second place has been underlined. See Section 4.2 for more details.

framework outperforms other models in all metrics except for ”Tissue,” which assesses the correctness of tissue reconstruction. This outcome aligns with our experimental observations, highlighting challenges in accurately reconstructing tissue boundaries, especially soft tissue, which tends to distort during the reconstruction process. Conversely, the RegGAN model showcases its proficiency in tissue preservation, suggesting that GAN-based approaches effectively mitigate variations during generation. Understanding the underlying mechanisms driving this phenomenon could potentially enhance our framework in the future. Further details of the experts' assessments can be found in supplementary materials.

4.2. Ablation Study

In our ablation studies, we systematically assess the effectiveness of the DiffuCE framework by sequentially removing its components, and we present the results in Table 2. Case studies are shown in supplementary materials.

Initially, we analyze by removing the DBE to evaluate the impact on domain bridging. Surprisingly, the performance doesn't decrease without the latent alignment module; instead, there is even a slight improvement in the metrics. However, a case study reveals that removing the DBE results in a reconstructed image with low contrast between soft tissue and the background.

Subsequently, we remove the CDD to assess the significance of the latent guidance modules. Without the CDD, the soft tissue representation is distorted, showing tissue-like patterns unrelated to the actual input CBCT image. Notably, the bone and lucent area are well-preserved, credited to the assistance from the CRD. The conditions are again utilized to guide the reconstruction of CT images.

Lastly, the CRD is removed to validate the efficacy of conditional refinement. The majority of metrics experience a significant drop compared to our baseline, highlighting the essential role of conditional refinement. Without it, the incorrect content produced by the CDD remains unchanged, resulting in poorer performance.

4.3. Limitation

Although our framework achieves the best performance in expert assessments, we acknowledge certain limitations in its current design. As mentioned in Sections 3.2 and 3.4, DiffuCE leverages conditional controls to influence the appearance of the output images, yielding satisfactory results. However, without conditional control, preserving intricate details becomes challenging, often resulting in distortions and lower performance on conventional pixel-level evaluations such as MAE and PSNR. We believe this limitation can be addressed by adding more condition-specific modules, such as those for muscle, soft tissue, or even tumors. Additionally, our framework's inference speed lags behind that of GAN-based algorithms, presenting a hurdle for deployment in real-world applications.

5. Conclusion

Addressing the critical clinical need for high-quality CT images while minimizing radiation dosage, we present DiffuCE, an efficient and effective framework designed to reconstruct detailed CT images from low-quality inputs.

To substantiate our framework's effectiveness, we conducted comprehensive validation. Our experiments encompassed quantitative analyses using pseudo-CBCT and CT datasets, complemented by experts' assessments employing the Likert scale. Moreover, the results on the public dataset show that our work has a competitive ability compared to SOTAs.

Furthermore, the introduction of latent alignment used in the encoder allows data from diverse domains to be mapped to the task domain using the respective alignment module, all while keeping the denoising UNet and its conditional guidance modules unaltered. This flexible design has the potential to evolve into more advanced frameworks for domain adaptation in latent diffusion models, making it a promising avenue for future research.

References

[1] Kancharagunta Kishan Babu and Shiv Ram Dubey. Cdgan: Cyclic discriminative generative adversarial networks for image-to-image transformation, 2021. 7

[2] Lianying Chao, Zhiwei Wang, Haobo Zhang, Wenting Xu, Peng Zhang, and Qiang Li. Sparse-view cone beam ct reconstruction using dual cnns in projection domain and image domain. *Neurocomputing*, pages 536–547, 2022. 2

[3] Qihang Chen, Zhidong Yuan, Chao Zhou, Weiguang Zhang, Mengxi Zhang, Yongfeng Yang, Dong Liang, Xin Liu, Hairong Zheng, Guanxun Cheng, and Zhanli Hu. Low-dose dental ct image enhancement using a multiscale feature sensing network. *Nuclear Inst. and Methods in Physics Research, A*, 981, 2020. 2

[4] Hyungjin Chung and Jong Chul Ye. Score-based diffusion models for accelerated mri. *Medical Image Analysis*, 80, August 2022. 1, 2

[5] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021. 2

[6] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. 7

[7] Yoseob Han and Jong Chul Ye. Framing u-net via deep convolutional framelets: Application to sparse-view ct. *IEEE Transactions on Medical Imaging*, 37:1418–1429, June 2018. 2

[8] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*, 2017. 6

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 1, 2, 3

[10] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *NIPS*, 2021. 2, 4

[11] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 3

[12] Evi M. C. Huijben, Maarten L. Terpstra, Arthur Jr. Galapon, Suraj Pai, Adrian Thummerer, Peter Koopmans, Manya Afonso, Maureen van Eijnatten, Oliver Gurney-Champion, Zeli Chen, Yiwen Zhang, Kaiyi Zheng, Chuanpu Li, Haowen Pang, Chuyang Ye, Runqi Wang, Tao Song, Fuxin Fan, Jingna Qiu, Yixing Huang, Juhyung Ha, Jong Sung Park, Alexandra Alain-Beaudoin, Silvain Bériault, Pengxin Yu, Hongbin Guo, Zhanyao Huang, Gengwan Li, Xueru Zhang, Yubo Fan, Han Liu, Bowen Xin, Aaron Nicolson, Lujia Zhong, Zhiwei Deng, Gustav Müller-Franzes, Firas Khader, Xia Li, Ye Zhang, Cédric Hémon, Valentin Boussot, Zhihao Zhang, Long Wang, Lu Bai, Shaobin Wang, Derk Mus, Bram Kooiman, Chelsea A. H. Sargeant, Edward G. A. Henderson, Satoshi Kondo, Satoshi Kasai, Reza Karimzadeh, Bulat Ibragimov, Thomas Helfer, Jessica Dafflon, Zijie Chen, Enpei Wang, Zoltan Perko, and Matteo Maspero. Generating synthetic computed tomography for radiotherapy: Synthrad2023 challenge report, 2024. 6, 7

[13] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 5

[14] Byeongjoon Kim, Hyunjung Shim, and Jongduk Baek. A streak artifact reduction algorithm in sparse-view ct using a self-supervised neural representation. In *Medical Physics*, 2022. 2

[15] Wei-An Lin, Haofu Liao, Cheng Peng, Xiaohang Sun, Jingdan Zhang, Jiebo Luo, Rama Chellappa, and Shaohua Kevin Zhou. Dudonet: Dual domain network for ct metal artifact reduction. *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[16] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps, 2022. 2

[17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1

[18] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 4, 6, 7

[19] Chenyu Shen, Ziyuan Yang, and Yi Zhang. Pet image denoising with score-based diffusion probabilistic models. In *MICCAI*, 2023. 1, 2

[20] J. Shtok, M. Elad, and M. Zibulevsky. Sparsity-based sinogram denoising for low-dose computed tomography. *IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP)*, 2011. 1, 2

[21] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022. 2

[22] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021. 1, 2

[23] Adrian Thummerer, Erik van der Bijl, Arthur Galapon, Joost J. C. Verhoeff, Johannes A. Langendijk, Stefan Both, Cornelis (Nico) A. T. van den Berg, and Matteo Maspero. Synthrad2023 grand challenge dataset: Generating synthetic ct for radiotherapy. *Medical Physics*, 50(7):4664–4674, June 2023. 6

[24] Jing Wang, Hongbing Lu, Tianfang Li, and Zhengrong Liang. Sinogram noise reduction for low-dose ct by statistics-based nonlinear filters. *Medical Imaging 2005: Image Processing*, Proc. SPIE 5747, 2005. 1, 2

[25] Yiying Wang, Tao Yang, and Weimin Huang. Limited-angle computed tomography reconstruction using combined fdk-based neural network and u-net. *IEEE Engineering in Medicine and Biology Society Conference Proceedings*, 2020. 2

[26] Yutong Xie and Quanzheng Li. Measurement-conditioned denoising diffusion probabilistic model for under-sampled medical image reconstruction. In *MICCAI*, 2022. 1, 2

[27] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *IEEE International Conference on Computer Vision (ICCV)*, 2023. 3

[28] Qinsheng Zhang and Yongxin Chen. Fast sampling of diffusion models with exponential integrator, 2023. 2

[29] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6

[30] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks, 2020. 7