bioengineering

MDPI

Article

Better Cone-Beam CT Artifact Correction via Spatial and Channel Reconstruction Convolution Based on Unsupervised Adversarial Diffusion Models

Guoya Dong [1,†], Yutong He [1,2,†], Xuan Liu [2], Jingjing Dai [2], Yaoqin Xie [2] and Xiaokun Liang [2,*]

[1] Hebei Key Laboratory of Bioelectromagnetics and Neural Engineering, School of Health Sciences and Biomedical Engineering, Hebei University of Technology, Tianjin 300130, China; dongguoya@hebut.edu.cn (G.D.); yt.he1@siat.ac.cn (Y.H.)
[2] Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China; xuan.liu@siat.ac.cn (X.L.); jj.dai@siat.ac.cn (J.D.); yq.xie@siat.ac.cn (Y.X.)
* Correspondence: xk.liang@siat.ac.cn
† These authors contributed equally to this work.

Abstract: Cone-Beam Computed Tomography (CBCT) holds significant clinical value in image-guided radiotherapy (IGRT). However, CBCT images of low-density soft tissues are often plagued with artifacts and noise, which can lead to missed diagnoses and misdiagnoses. We propose a new unsupervised CBCT image artifact correction algorithm, named Spatial Convolution Diffusion (ScDiff), based on a conditional diffusion model, which combines the unsupervised learning ability of generative adaptive networks (GAN) with the stable training characteristics of diffusion models. This approach can efficiently and stably achieve CBCT image artifact correction, resulting in clear, realistic CBCT images with complete anatomical structures. The proposed model can effectively improve the image quality of CBCT. The obtained results can reduce artifacts while preserving the anatomical structure of CBCT images. We compared the proposed method with several GAN- and diffusion-based methods. Our method achieved the highest corrected image quality and the best evaluation metrics.

Keywords: CBCT reconstruction; diffusion model; deep learning

check for updates

Academic Editor: Firstname Lastname

Received: 11 December 2024
Revised: 5 January 2025
Accepted: 14 January 2025
Published: 30 January 2025

Citation: Dong, G.; He, Y.; Liu, X.; Dai, J.; Xie, Y. and Liang, X. Better Cone-Beam CT Artifact Correction via Spatial and Channel Reconstruction Convolution Based on Unsupervised Adversarial Diffusion Models. *Bioengineering* **2025**, *12*, 132. https://doi.org/10.3390/bioengineering12020132

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Cone-Beam Computed Tomography (CBCT) is a medical imaging technology that uses cone-beam X-ray scanners and digital imaging techniques to obtain three-dimensional images. Compared to planning CT (pCT), CBCT offers greater real-time capabilities, shorter scan times, and lower X-ray doses. Therefore, CBCT holds significant clinical value in image-guided radiotherapy (IGRT). However, low-density soft tissues in CBCT images are often plagued with artifacts and noise, which can lead to missed diagnoses and misdiagnoses.

To reduce the impact of CBCT image artifacts on clinical diagnosis and improve diagnostic accuracy, researchers have proposed various methods. These methods can be divided into hardware-based correction methods and software-based correction. Due to the high cost and complex operation of hardware-based correction methods, software-based correction methods are preferred by researchers. In recent years, deep learning, which has developed rapidly in medical image processing [1,2], can be applied to image denoising [3], sparse reconstruction [4], artifact correction [5], and so on. Many deep learning-based methods have also been proposed in CBCT image artifact correction. Kida S et al. [6] applied U-Net to improve the spatial uniformity of CBCT images. Chang et al. [7] applied

CNN to generate images with fewer artifacts directly from the sinogram domain, which can suppress the ring artifact effectively without the introduction of structure distortion.

However, the training of these deep learning methods often relies on a large number of paired data with or without artifacts, which is difficult to obtain in clinical scenarios. Recently, generative adversarial networks (GANs) [8] have made rapid progress in the field of image generation [9], style migration [10], and data augmentation [11]. Researchers try to introduce the unsupervised GAN-based methods into the CBCT image artifact correction task. Liang et al. [12] utilized a generative adversarial network framework with cycle-consistency (CycleGAN), which is capable of using unpaired CT and CBCT images to achieve image-to-image translation in an unsupervised manner. Dong et al. [13] used a multilayer and patch-based method to translate the low-quality CBCT images to high-quality CBCT images. Wang et al. [14] combined a double contrast learning adversarial network framework (DCLGAN) and post-processing techniques to obtain high-quality CBCT images. Image restoration methods based on GANs effectively reduce artifacts in CBCT images [12], and they can capture correlations in the data without requiring precisely labeled training data. This solves the problem of obtaining precise labels for medical images. However, the learning process of GANs may suffer from mode collapse, making them difficult to train [15]. In addition, images generated by GANs may have unreasonable or discontinuous parts in the overall structure, resulting in significant differences from the original images. Compared with other generative networks, GANs are over flexible, which makes simple GANs less controllable for larger images with more pixels.

To generate high-quality images and make the generation process more controllable, researchers have introduced diffusion model [16] into medical image restoration tasks. Ozbey M et al. [17] proposed an adversarial diffusion model named SynDiff to translate images between MRI and CT modalities. However, the network structure of SynDiff is complex, and this method will produce a gradient explosion in the training process. Li et al. [18] have proposed a Frequency-domain Guided Diffusion Model (FGDM) for image translation. However, this method only uses high-quality images and is based on an empirical frequency domain assumption without utilizing information from unpaired data.

This paper addresses the problems of artifact correction in CBCT images, with the following challenges: (i) It is difficult to obtain precisely labeled medical image training data; (ii) Most existing GAN- or diffusion-based methods require extensive computational resources, time, and large training datasets. To tackle the above challenges, we designed a novel CBCT image artifact correction method named the Spatial Convolution Diffusion (ScDiff) model. To solve the problem of obtaining accurate labels for medical images, the Match Module in ScDiff generates fake CBCT images based on cyclic-consistency loss, providing a reference for the Diffusive Module. The Diffusive Module is guided by the generated fake CBCT images and utilizes a large-step conditional diffusion process for efficient and accurate image sampling. In order to solve the problem of low training efficiency of diffusion models, we introduce the Spatial and Channel Reconstruction Convolution (SCConv) module [19] to the downsampling process of the Diffusive Module. This module can simultaneously process the spatial (shape, structure) and channel (depth) information of the image, making the network more refined and efficient in CBCT images generating.

We validated our network on the classical public TCIA lung datasets [20–24]. Compared to the most advanced method FGDM model, ScDiff showed significant improvements in metrics such as MAE, RMSE, PSNR, and SSIM.

In summary, the contributions of this paper are as follows:

(i) ScDiff combines the advantages of the unsupervised learning of GANs and the stable training characteristics of diffusion models. It enables stable and efficient training and synthesizing more realistic and clearer synthetic CBCT (sCBCT) images.

(ii) By introducing the SCConv module into ScDiff, the network training efficiency is effectively enhanced by reducing feature redundancy in network analysis.

(iii) We compared ScDiff with the SOTA method, Frequency-domain Guided Diffusion Model (FGDM). The results indicated that our method performs the best on all indicators.

2. Materials and Methods

2.1. Diffusion Model

The diffusion model transforms the original image into pure Gaussian noise by gradually adding noise in the forward process. In the reverse process, the diffusion model restores high-quality images from pure noise. By training the model to approximate the reverse process, we can generate complex data distributions from a simple noise distribution.

Given a clean image x_0, the forward diffusion process gradually adds Gaussian noise to the input image to form samples on the timesteps. By adding Gaussian noise gradually on the basis of the previous step x_{t-1}, the current time x_t is obtained. This process can be regarded as a Markov chain. The transition probability of the state at the next moment depends only on its previous state. From x_{t-1} to x_t, the mapping is as follows:

$$x_t = \sqrt{1-\theta_t}x_{t-1} + \sqrt{\theta_t}\epsilon, \epsilon \sim \mathcal{N}(0, I) \tag{1}$$

where t is the time step, $t \in \{0, 1, 2...T\}$. θ_t is the variance of the added noise for each step, which is an increasing sequence, ϵ is the added noise, which follows a unit Gaussian distribution, and x_t is the data distribution of step t. The corresponding forward transition probability is as follows:

$$q(x_t \mid x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\theta_t}x_{t-1}, \theta_t I\right)(t = 1, 2, \ldots, T) \tag{2}$$

where $q(x_t \mid x_{t-1})$ is the forward transition probability, $\mathcal{N}$ is the Gaussian distribution and I is the unit diagonal matrix. By repeatedly adding noise, any input image x_0 can be converted into a sample x_T close to unit Gaussian noise.

In the reverse process, the task of the diffusion model is to remove the added noise and restore the original input image from a pure noise image. In the case of large T and small θ_t, the posterior probabilities of x_{t-1} and x_t can be obtained from the Bayesian formula, which is an approximately Gaussian distribution:

$$p(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1}; \mu(x_t, t), \Sigma(x_t, t)) \tag{3}$$

where $p(x_{t-1} \mid x_t)$ is the reverse transition probability, $\mu(x_t, t)$ is the mean and $\sum(x_t, t)$ is the variance. They are both predicted by the model.

Usually, the variational lower bound (VLB) is used to train the diffusion model. The training goal is to minimize the KL divergence between the data distribution and the generation distribution. After model training, the generation process can gradually generate data from noise in the following ways:

$$x_{t-1} = \frac{1}{\sqrt{\gamma_t}}\left(x_t - \frac{\theta_t}{\sqrt{1-\bar{\gamma}_t}}\epsilon(x_t, t)\right)(t = T, T-1, \ldots, 1) \tag{4}$$

where $\gamma_t = 1 - \theta_t$.

2.2. ScDiff

To improve the quality of CBCT, we propose an unsupervised conditional diffusion-based model. The network architecture is shown in Figure 1. The ScDiff model includes the Match Module and the Diffusive Module. The Match Module is used to generate fake CBCT

images paired with CT images to realize unsupervised learning without the requirements of paired data.

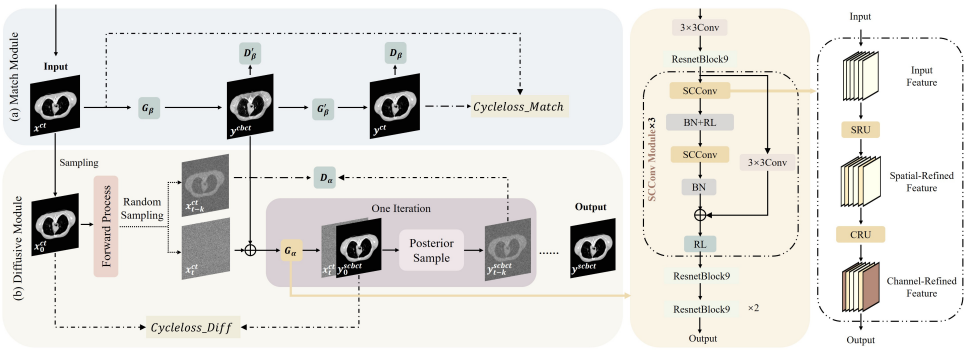

Figure 1. An overview of the ScDiff framework: (a) Match Module. (b) Diffusive Module. Each purple block shows one iteration for calculating $\hat{y}^{scbct}_{t-k}$ from x^{ct}_t while x^{ct}_t is sampled from a unit Gaussian distribution. The right part of the figure shows the details of our proposed SCConv Module.

The Diffusive Module improves the forward process of a traditional diffusion model. The traditional diffusion model usually needs a large time t to ensure that the step is small enough to meet the normality assumption [25]. This method will limit the efficiency of image generation. Specifically, we use a large step k in the Diffusive Module to achieve an efficient sampling process. In the reverse process of the Diffusive Module, the generator first denoising the current noisy image to obtain a prediction as close to the clean image as possible.

Match Module Given unpaired CBCT and pCT images x^{cbct} and x^{ct}. We obtain x^{cbct}_0 and x^{ct}_0 by random sampling. Then, we use a generator G_β to translate pCT images to fake CBCT images and other G'_β to translate fake CBCT images back to pCT images. The generated fake CBCT images are denoted as $\hat{y}^{cbct}$:

$$\hat{y}^{cbct}_0 = G_\beta(x^{ct}_0), \quad \hat{y}^{ct}_0 = G'_\beta\left(\hat{y}^{cbct}_0\right), \tag{5}$$

Two discriminators D_β and D'_β are used for judging the authenticity of the generated $\hat{y}^{cbct}_0$ and $\hat{y}^{ct}_0$. For G_β, D_β and G'_β, D'_β, unsaturated adversarial loss [26] is adopted:

$$\begin{aligned} L_{G_\beta} &= \mathbb{E}_{p_\beta(x^{cbct}_0|x^{ct}_0)}\left[-\log\left(D_\beta(\hat{y}^{cbct}_0)\right)\right] \\ L_{G'_\beta} &= \mathbb{E}_{p'_\beta(x^{ct}_0|\hat{y}^{cbct}_0)}\left[-\log\left(D'_\beta(\hat{y}^{ct}_0)\right)\right] \end{aligned} \tag{6}$$

$$\begin{aligned} L_{D_\beta} &= \mathbb{E}_{q(x^{cbct}_0|x^{ct}_0)}\left[-\log\left(D_\beta(x^{cbct}_0)\right)\right] + \mathbb{E}_{p_\beta(x^{cbct}_0|x^{ct}_0)}\left[-\log\left(1-D_\beta(\hat{y}^{cbct}_0)\right)\right] \\ L_{D'_\beta} &= \mathbb{E}_{q(x^{ct}_0|\hat{y}^{cbct}_0)}\left[-\log\left(D'_\beta(x^{ct}_0)\right)\right] + \mathbb{E}_{p'_\beta(\hat{y}^{ct}_0|x^{cbct}_0)}\left[-\log\left(1-D'_\beta(\hat{y}^{ct}_0)\right)\right] \end{aligned} \tag{7}$$

where $p_\beta\left(x^{cbct}_0 \mid x^{ct}_0\right)$ and $p'_\beta\left(\hat{y}^{ct} \mid x^{cbct}_0\right)$ mean the network parametrization of the conditional distribution of x^{cbct}_0 and $\hat{y}^{ct}$ given the x^{ct}_0 and x^{cbct}_0 image. $q\left(x^{cbct}_0 \mid x^{ct}_0\right)$ and $q\left(x^{ct}_0 \mid \hat{y}^{cbct}_0\right)$ represent the true conditional distribution of the image obtained by the generator.

In order to ensure the consistency between the fake CBCT images and pCT images, we use the cycle-consistency loss to constrain the performance of the model. Comparing the images $\hat{y}^{ct}_0$ generated by the Match Module with the input x^{ct}_0, the cycle-consistency loss function is obtained:

$$L_{\text{cycM}} = \mathbb{E}_{t,q(x^{ct}_0)}(\lambda_1|x^{ct}_0 - \hat{y}^{ct}_0|_1) \tag{8}$$

where $\hat{y}_0^{ct}$ is obtained from $G'_\beta(\hat{y}_0^{cbct})$. λ_1 is the weight of the cycle-consistency loss item of the Match Module. The ℓ_1-norm of the difference between two images is used as a consistency measure [27].

The loss of the Match Module is as follows: $L_{Match} = L_{G_\beta} + L_{G'_\beta} + L_{D_\beta} + L_{D'_\beta}$

Diffusive Module In this module, we used the fake CBCT images generated by the Match Module as a condition to predict its corresponding sCBCT images. In the forward process of the Diffusive Module, we use k as a time step and add Gaussian noise to x_0^{ct} step by step until T. We can obtain different $\{x_0^{ct}, x_k^{ct} ... x_{t-k}^{ct}, x_t^{ct} ... x_T^{ct}\}$ with different levels of noise. Then we input $(x_t^{ct}, \hat{y}_0^{cbct})$ and current time $t \sim \mathcal{U}(\{0, k, \ldots, T\})$ to generator $G_\alpha(x_t^{ct}, \hat{y}_0^{cbct}, t)$. Each k step generates a deterministic estimate $\hat{y}_{t-k}^{scbct}$ of the pCT image. After the iterations, we obtain $\hat{y}_0^{scbct}$. The transition probability is $q\left(x_{t-k}^{ct} \mid x_t^{ct}, \hat{y}_0^{cbct}\right)$. After the discrimination of the discriminator D_α, the final result closest to the real noiseless target image is obtained. G_α uses unsaturated counter loss [26], D_α uses unsaturated counter loss with gradient penalty [28]:

$$L_{G_\alpha} = \mathbb{E}_{t, q\left(x_t^{ct} | x_0^{ct}, \hat{y}_0^{cbct}\right), p_\alpha\left(x_{t-k}^{ct} | x_t^{ct}, \hat{y}_0^{cbct}\right)} \left[-\log\left(D_\alpha\left(\hat{y}_{t-k}^{scbct}\right)\right)\right] \tag{9}$$

$$\begin{aligned} L_{D_\alpha} = \mathbb{E}_{t, q\left(x_t^{ct} | x_0^{ct}, \hat{y}_0^{cbct}\right)} \Big[& \mathbb{E}_{q\left(x_{t-k}^{ct} | x_t^{ct}, \hat{y}_0^{cbct}\right)} \left[-\log\left(D_\alpha\left(x_{t-k}^{ct}\right)\right)\right] \\ & + \mathbb{E}_{p_\alpha\left(x_{t-k}^{ct} | x_t^{ct}, \hat{y}_0^{cbct}\right)} \left[-\log\left(1 - D_\alpha\left(\hat{x}_{t-k}^{scbct}\right)\right)\right] \\ & + \eta \mathbb{E}_{q\left(x_{t-k}^{ct} | x_t^{ct}, \hat{y}_0^{cbct}\right)} \left\| \nabla_{x_{t-k}^{ct}} D_\alpha\left(x_{t-k}^{ct}\right) \right\|_2^2 \Big] \end{aligned} \tag{10}$$

where η is the weight of the gradient penalty.

In order to ensure the consistency between the sCBCT images generated by the Diffusive Module and pCT images, we use the cycle-consistency loss to constrain the performance of the Diffusive Module. Comparing the images $\hat{y}_0^{scbct}$ with the x_0^{ct} in the end of every step, the cycle-consistency loss function is obtained:

$$L_{\text{cycD}} = \mathbb{E}_{t, q(x_t^{ct} | x_0^{ct})} (\lambda_2 | x_0^{ct} - \hat{y}_0^{scbct} |_1) \tag{11}$$

where $\hat{y}_0^{scbct}$ is obtained from $G_\alpha(x_t^{ct}, \hat{y}_0^{cbct}, t)$. λ_2 is the weight of the cycle-consistency loss item of the Match Module. The ℓ_1-norm of the difference between two images is used as a consistency measure [27].

The loss of Diffusive Module is $L_{Diff} = L_{G_\alpha} + L_{D_\alpha}$, and the cycle-consistency loss of the ScDiff is $L_{cyc} = L_{\text{cycM}} + L_{\text{cycD}}$

The Match Module and Diffusive Module are trained jointly. The total loss function of the ScDiff is as follows:

$$L_{\text{total}} = \lambda_3 L_{Match} + \lambda_4 L_{Diff} + L_{\text{cyc}} \tag{12}$$

where λ_3, λ_4 are the weight of the adversarial loss term of the Match Module and the Diffusion Module. The Match Module provides paired predictive images for the Diffusion Module during the training process. During inference process, Diffusion Module only needs to execute the generator G_α. Starting from time T, the generator G_α gradually obtains the target image step by step and uses the result of the previous step as an input sample for the next step. Finally, output clean sCBCT images $\hat{y}^{scbct}$.

SCConv Module In order to reduce feature redundancy and save computational costs, we introduce SCConv [19] in the downsampling process of generator G_α as shown on the right side of Figure 1. This module consists of two parts, Spatial Reconstruction Unit (SRU) and Channel Reconstruction Unit (CRU). When inputting $(x_t^{ct}, \hat{y}_0^{cbct})$, the feature f_0 is

obtained through a 3×3 convolution and a ResnetBlock module. SRU first performs group normalization (GN) on feature f_0:

$$f = GN(f_0) = \gamma \frac{f_0 - \mu}{\sqrt{\sigma^2 + \varepsilon}} + z \tag{13}$$

where μ and σ are the mean and standard deviation of f_0. ε is a small positive number added for stable division, while γ and z are trainable affine transformations. γ is used to measure the spatial pixel variance of each batch and channel. The larger the γ, the more spatial pixels change, and the richer the spatial information. We can obtain the normalized correlation weight W_γ representing the importance of different feature maps through Equation (14). Mapping W_γ to the range $(0, 1)$ using the sigmoid function and gate it with a threshold of 0.5.

$$W_\gamma = \{w_i\} = \frac{\gamma_i}{\sum_{j=1}^{C} \gamma_j}, \quad i, j = 1, 2, \cdots, C \tag{14}$$

where C is the channel of f. The weight greater than the threshold is set to 1 to obtain the information weight W_1, and the weight less than the threshold is set to 0 to obtain the redundant weight W_2. Multiply the information weight and redundant weight with f element by element to obtain f_1^w, which has informative and expressive spatial contents, and f_2^w, which has little or no information:

$$\begin{aligned} f_1^w &= W_1 \otimes f \\ f_2^w &= W_2 \otimes f \end{aligned} \tag{15}$$

where $\otimes$ represents element-wise multiplication. In order to fully combine the weighted two different information rich features and enhance the information flow between them, cross reconstruction [29] is used to obtain f^{w_1} and f^{w_2}. Finally, connect f^{w_1} and f^{w_2} to obtain f^w.

$$f^w = f^{w_1} \cup f^{w_2} \tag{16}$$

The representative features in f are enhanced in f^w, and the redundancy of f in the spatial dimension is suppressed. However, there is still channel redundancy in f^w. So, CRU first performs a split operation to handle the channel redundancy in f^w. Divide f^w into vC with v channel and $(1 - v)C$ with $1 - v$ channels. Channel compression is performed through 1×1 convolution to obtain f_{vC}^w and $f_{(1-v)C}^w$.

Next, use a 3×3 groupwise convolution (GWC) [30] with g = 2 and a 1×1 pointwise convolution (PWC) [31] to extract information from f_{vC}^w. Add the output results together to form Y_1^C. Extract shallow details from $f_{(1-v)C}^w$ using 1×1 PWC as a supplement to f_{vC}^w. Connect the output of PWC with $f_{(1-v)C}^w$ to obtain Y_2^C. Finally, we adaptively merge Y_1^c and Y_2^c using a simplified SKNet method [32] to obtain the final output Y.

The SCConv module reduces the spatial and channel redundancy of input features through SRU and CRU, effectively reducing computational costs and improving model performance.

3. Experiments

3.1. Experimental Settings

Datasets. The data collection is made up of images from 20 patients with locally advanced non-small cell lung cancer taken throughout chemotherapy and radiation therapy. These images are sourced from TCIA [20–24], with undergoing CBCT and pCT scans for each patient. After registering and filtering out poor-quality data, a total of 6721 slices of CBCT and pCT images were obtained. For training, 16 patients contributed 5377 slices. Four

patients provided 1344 slices for testing. The axial matrix size of CBCT and pCT images is 512×512. All images in the training dataset were normalized to the range $[-1, 1]$.

Training and Inference. This model is trained by using an NVIDIA RTX 3090 with 24 GB of RAM. Training was carried out using the Adam optimization technique with $\beta 1$ set to 0.5 and $\beta 2$ set to 0.9, as these parameter values have been demonstrated to work well in similar deep learning tasks and can help the model converge stably during the training process. On the test set in each dataset, the model's performance was assessed. After finishing the network training, the sCBCT images were produced. The model took 360 h for training. Meanwhile, during training, it took about 30 min for the diffusion model to conduct image production training for each test patient round.

Evaluation Metric. For quantitative assessment, lung cancer patient image pairs from CBCT and pCT were utilized. The reference was the pCT. A comparison of the axial views of pCT and CBCT images was made at the same window level in order to confirm that our strategy has improved the quality of synthetic pCT images. The structural similarity and spatial uniformity of the generated images, as well as the improvement of sCBCT over CBCT, were statistically evaluated using the mean absolute error (MAE) [33], root mean square error (RMSE) [33], peak signal-to-noise ratio (PSNR) [34] and structural similarity index (SSIM) [35]. The following are the definitions of these metrics between sCBCT and pCT:

$$\begin{aligned} MAE &= \frac{1}{M} \sum_{i,j}^{n_{sCBCT} n_{pCT}} |sCBCT(i,j) - pCT(i,j)|, \\ RMSE &= \sqrt{\frac{1}{n_{sCBCT} n_{pCT}} \sum_{i,j}^{n_{sCBCT} n_{pCT}} (sCBCT(i,j) - pCT(i,j))^2}, \\ PSNR &= 10 \log_{10} \left(\frac{MAX^2}{\sum_{i,j}^{n_{sCBCT} n_{pCT}} \frac{(sCBCT(i,j) - pCT(i,j))^2}{n_{sCBCT} n_{pCT}}} \right), \\ SSIM &= \frac{(2\mu_{sCBCT}\mu_{pCT} + c_1)(2\sigma_{sCBCT \cdot pCT} + c_2)}{(\mu_{sCBCT}^2 + \mu_{pCT}^2 + c_1)(\sigma_{sCBCT}^2 + \sigma_{pCT}^2 + c_2)}. \end{aligned} \tag{17}$$

Competing Methods. We compared several SOTA GAN - and diffusion-based methods with ScDiff. All competing methods use unpaired CBCT and pCT for unsupervised learning. We adjusted the hyperparameters of each method to improve the performance of the validation set. The adjusted parameters include epoch, learning rate, and loss weight. The hyperparameters of ScDiff were 50 epochs, 2×10^{-5} learning rate, T = 1000, a step size of k = 250, and diffusion steps T/k = 4. Weights for cycle-consistency were λ_3, λ_4 = 0.5. The hyperparameters of CycleGAN, CUT, and DCLGAN were 100 epochs, a 10^{-4} learning rate linearly decayed to 0 in the last 50 epochs. The hyperparameters of SynDiff were 50 epochs, 10^{-4} learning rate, T = 1000, a step size of k = 250, and diffusion steps T/k = 4. Weights for losses were $\lambda_{1\phi,1\theta}$ = 0.5 and $\lambda_{2\phi,2\theta}$ = 1. The hyperparameters of FGDM were 50 epochs, 10^{-4} learning rate, T = 1000, k = 1, and 1000 diffusion steps. Weight for loss was 1.

3.2. Results

Figure 2 illustrates the axial, sagittal, and coronal views of two patients' CBCT, sCBCT, and pCT images. It can be observed that the stripe artifacts in the CBCT image are severe, leading to partial tissue loss and significant CT value variations. In comparison to CBCT, the generated sCBCT image demonstrates effective artifact suppression and noticeable improvements in soft tissue contrast, spatial uniformity, and clarity.

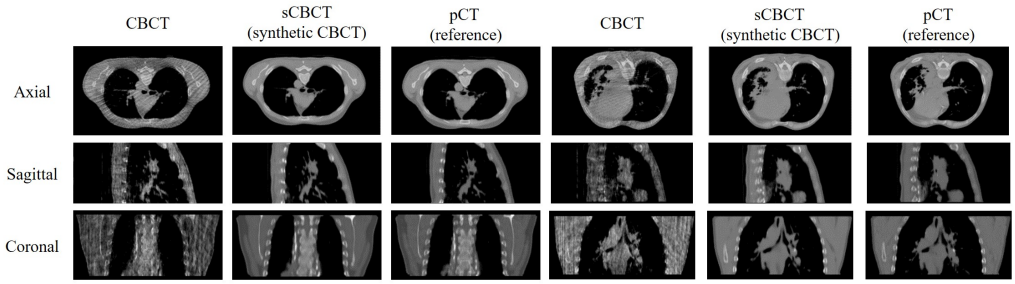

Figure 2. Comparisons of the image quality between CBCT, sCBCT (proposed method), and pCT (reference) during a particular patient phase. The images are sagittal, coronal, and axial in the top, middle, and bottom rows, respectively. The CBCT, sCBCT (proposed method), and pCT (reference) are indicated by the left, center, and right, respectively.

Furthermore, Figure 3 presents the axial view of a selected patient, with the regions of interest (ROI) region outlined in red and green. It is evident that the quality of the sCBCT is significantly superior to the CBCT image, with image quality and spatial uniformity being enhanced while anatomical consistency is maintained. The third row focuses on skeletal structures, and despite registration issues resulting in slight anatomical discrepancies between CBCT and pCT, the anatomical consistency between sCBCT and CBCT images indicates faithful restoration of CBCT's anatomical information. The violin plot shows that the sCBCT image is closer to the pCT image in terms of HU values, indicating that the proposed method produces more realistic images.

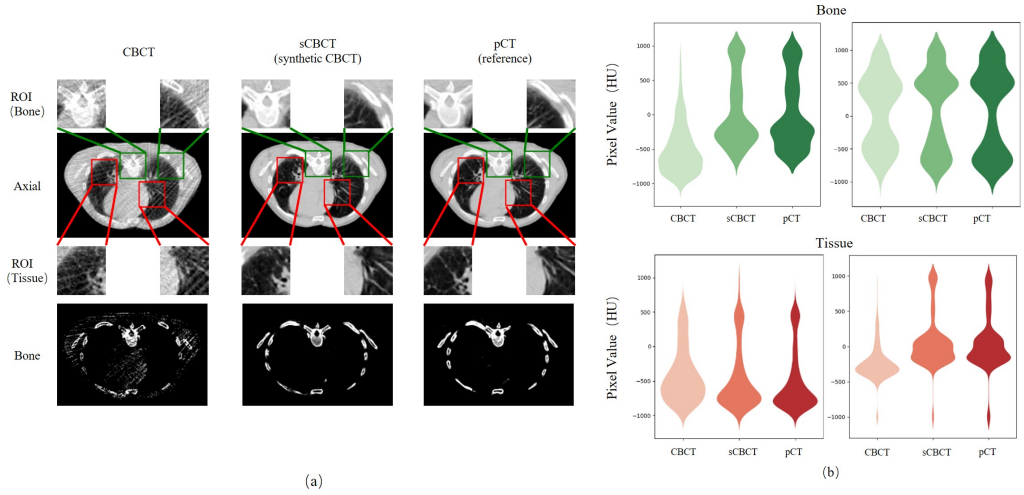

Figure 3. Comparison of HU values in ROI. (**a**) Red-colored boxes in the images are zoomed to demonstrate the tissue and green-colored boxes in the images are zoomed to demonstrate the bone. The range of the CT number display window is [−1000, 1000] HU. The fourth row simply shows the bone structure, excluding the soft tissue. The window for display is [500,750] HU. (**b**) Comparison of HU values in ROI with Violin diagram.

The distribution of HU values is illustrated in Figure 4, with the red line representing the longitudinal distribution and the yellow line representing the transverse distribution. The results show that, compared to CBCT, sCBCT exhibits HU values that are closer to those of pCT, indicating that sCBCT can not only reduce artifacts but also correct HU value discrepancies.

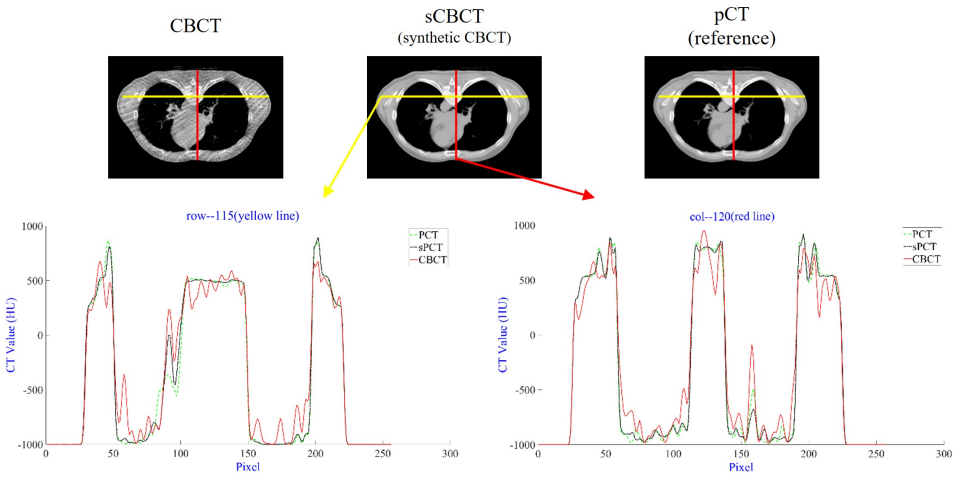

Figure 4. In the first row, CBCT, sCBCT, and pCT are displayed in axial perspectives from left to right. The HU value distributions for CBCT, sCBCT, and pCT are shown in the second row along yellow lines (line 115) and red lines (line 120). The CBCT, sCBCT, and pCT HU value distributions are shown by the red, black, and green lines, respectively. The display window's current setting is [−1000, 1000] HU.

Figure 5 and Table 2 illustrate the qualitative comparison and quantitative comparison of the results obtained by different methods, respectively. In terms of visual effects, all models can achieve the elimination of artifacts compared to CBCT. Specifically, the images generated by CycleGAN and CUT seem to be satisfactory. However, as can be seen from the zoomed-in green box, these two methods cannot completely eliminate stripe artifacts and will result in structural losses, such as soft tissue and spinal regions. Additionally, there is a noticeable blurring of the boundaries of soft tissue and a significant CT value error. FGDM exhibits a less effective image restoration under the same number of iterations and hyperparameters, particularly for high-contrast anatomical features like the skeleton. Relatively, images generated by DCLGAN and SynDiff better preserve anatomical details, striking a balance between realism and fidelity, although they still fall short compared to ScDiff. As highlighted in the green box, ScDiff eliminates stripe artifacts and effectively retains structural details compared to other methods. Even though ScDiff works the best out of all the techniques, there are still a few small structural inconsistencies, with the source images needing to be worked out and optimized. From the difference plot on the right, the difference between ScDiff and pCT is the smallest. This indicates that the proposed method can effectively reduce the difference between CBCT and pCT. The obtained results can preserve more anatomical structures and have a better effect on removing artifacts.

The quantitative evaluation results are presented in Table 1, along with the average values. This covers the CBCT, sCBCT, and pCT compared to pCT MAE, RMSE, PSNR, and SSIM values. Spatial consistency and structural similarity around pCT are indicated by sCBCT's MAE value of 19.398 HU and RMSE value of 62.707 HU, respectively, compared with CBCT. After correction, PSNR increases from 24.540 dB to 30.469 dB, and SSIM increases from 0.811 to 0.924, suggesting that the generated sCBCT images are of lower distortion and better similarity to pCT in terms of brightness, contrast, and structure, demonstrating the effectiveness of our denoising approach.

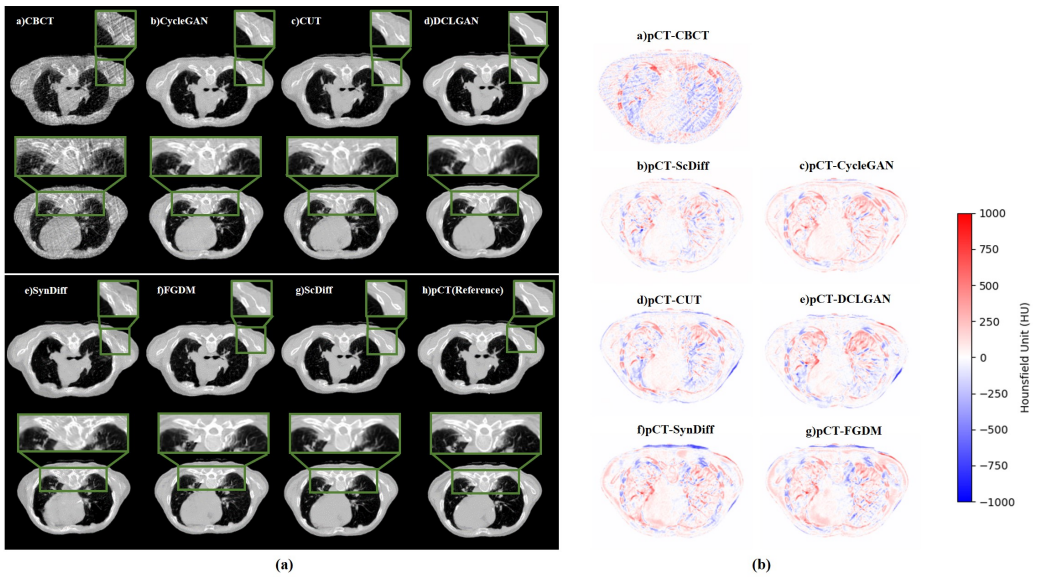

Figure 5. (**a**) Comparison of the image quality produced by using various methods. The inference sample is the pCT. The green box displays local zoomed-in images of the results obtained by different methods. The display window is [−1000, 1000] HU. (**b**) Difference images between pCT and different methods.

Table 1. Comparison results for each patient in test datasets with MAE (HU), RMSE (HU), PSNR (dB), and SSIM.

Data	Image Types	MAE↓	RMSE↓	PSNR↑	SSIM↑
Patient 1	sCBCT-pCT	18.496	60.093	30.662	0.926
Patient 2	sCBCT-pCT	18.319	59.662	30.753	0.927
Patient 3	sCBCT-pCT	20.803	67.172	30.263	0.921
Patient 4	sCBCT-pCT	19.852	63.899	30.196	0.923
Mean	sCBCT-pCT	**19.398**	**62.707**	**30.469**	**0.924**
	CBCT-pCT	49.945	117.803	24.540	0.811

Remark: "sCBCT" denotes synthetic CBCT after correction. ↑ (↓) indicates that the larger (smaller) the value, the better the performance; the **Bold** numbers indicate this metric's best performance.

As demonstrated in Table 2, in a quantitative comparison of CycleGAN, CUT, DCLGAN, SynDiff, and FGDM, CycleGAN and DCLGAN perform less well in RMSE but comparatively better in SSIM. In general, additional loss functions used by CycleGAN, CUT, and DCLGAN to preserve structure have a negative impact on image quality and lower PSNR. Although SynDiff has shown overall good results in CBCT artifact correction tasks, it still lags behind ScDiff in PSNR and SSIM. FGDM, another diffusion model-based approach, has a low SSIM of only 0.86, indicating poor performance. All things considered, ScDiff performs better than other techniques on every metric. This is attributed to its capacity to inherit both the realistic and high-quality image generation ability of the diffusion model and the image preservation capabilities of GAN.

Table 2. Quantitative comparison of the image quality produced using various techniques.

	Methods	MAE↓	RMSE↓	PSNR↑	SSIM↑	Inference Time↓
GAN-Based	CycleGAN [12]	36.497	97.483	26.510	0.853	**0.236**
	CUT [13]	37.764	101.219	25.991	0.850	0.258
	DCLGAN [14]	32.491	<u>88.273</u>	27.227	0.876	0.278
Diffusion-Based	SynDiff [17]	34.845	94.521	26.730	0.862	0.312
	FGDM [36]	<u>31.170</u>	88.825	<u>27.323</u>	<u>0.881</u>	0.397
	ScDiff	**19.398**	**62.707**	**30.469**	**0.924**	**0.228**

Remark: ↑ (↓) indicates that the larger (smaller) the value, the better the performance; the **Bold** and <u>underlined</u> numbers indicate this metric's best and second-best performance, respectively.

After the correction of ScDiff, the quality of CBCT images has significantly improved, with clearer anatomical structures and higher contrast. Doctors can observe and evaluate lesion areas more accurately through corrected sCBCT images and develop more precise treatment plans. The corrected sCBCT image plays a crucial role in image-guided radiotherapy (IGRT). It can significantly improve the accuracy of positioning and target area delineation, laying the foundation for precision radiotherapy.

3.3. Ablation Study

In ScDiff, cycle-consistency loss plays a role in improving generated image quality and ensuring generated image consistency. Meanwhile, ScDiff adopts the SCConv module to address the limitations of traditional convolution operations in both spatial and channel dimensions [19]. To test their contributions, we conducted a study of ScDiff through the ablation of different modules in this sub-section. The results of the ablation study are shown in Table 3. It can be seen that the introduction of cycle-consistency loss maintains the ability of high-quality image generation, while the SCConv module improves the efficiency of network training and testing.

Table 3. Average inference time per slice and quantitative comparison of the image quality.

Baseline	Cycle-Consistency Loss	SCConv	Perceptual Loss	Inference Time (s) ↓	PSNR↑	SSIM↑
√				0.312	24.412	0.736
√	√			0.342	<u>29.330</u>	<u>0.862</u>
√		√		**0.194**	24.711	0.753
√	√	√		<u>0.228</u>	**29.976**	**0.919**
√	√	√	√	0.237	20.853	0.615

Remark: ↑ (↓) indicates that the larger (smaller) the value, the better the performance; the **Bold** and <u>underlined</u> numbers indicate this metric's best and second-best performance, respectively.

4. Discussion

We proposed an unsupervised CBCT artifact correction method based on conditional diffusion, ScDiff, which includes the Match Module and Diffusive Module. The Match Module introduces cycle-consistency loss in two generator-projector pairs to generate CBCT images paired with the target sCBCT images. The Diffusive Module receives the output of the Match Module as a guide and uses the CT image as the input to generate the sCBCT image. The forward process adds noise to the CT image. In the inverse process of diffusion, the generator first denoises the current noisy image to obtain a prediction as close to the clean image as possible.

The task of this paper is essentially a medical image translation task, and the size and specificity of the dataset used for training are limited. The output of the model depends on the CBCT image in anatomical structure. This requires the Match Module to output

excellent CBCT images. This conclusion is verified in the experiment. If the output obtained from the Match Module shows underfitting during training and is used for subsequent steps, the final result cannot achieve the best effect. In addition, we found that gradient explosion occurred in the Diffusive Module during the training process. This is due to the high learning rate. And the scale of noise increases or decreases gradually in the process of forward and reverse denoising of the diffusion model. If the scale of noise is not adjusted properly in this process, it may lead to too large a gradient at high time steps, which makes it difficult to train the model stably. In order to solve this problem, we compared the training loss of the ScDiff model at different learning rates and the training loss and validation loss of ScDiff under different dropout rates. As shown in Figure 6. When lr is set to 1e-3, a gradient explosion phenomenon occurs during the model training process. When lr is set to 1e-4, the model requires more epochs for convergence. So, we ultimately used lr = 5e-4 as the parameters for model training. However, overfitting occurred during the validation process, as shown in the first column of Figure 6b when Dropout = 0. To address the issue of overfitting, we set the dropout values to 0.2 and 0.5, respectively. The results show that the model can fit normally when dropout = 0.2.

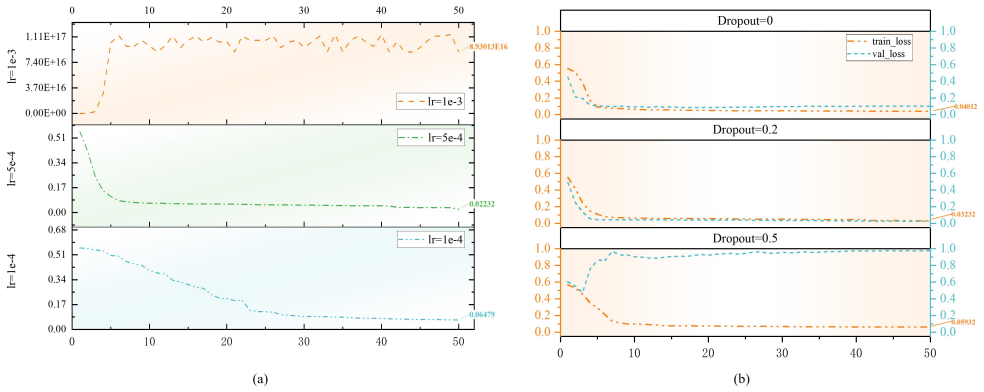

Figure 6. (**a**) Training loss curves of ScDiff model at different learning rates. (**b**) Training and validation loss curves of ScDiff model at different dropout rates.

Although the proposed ScDiff has shown encouraging performance in correcting scattering artifacts in CBCT, there are still some problems to be solved. When removing fringe artifacts, a small number of small lung textures in the feature image are ignored. This may lead to missed diagnosis or misdiagnosis. In addition, a CBCT scanner with a shorter scanning time may be used in clinical. The singing of the breathing cycle and the shortening of the projection interval will lead to a further decline in image quality. The current experiment only considers a single scanning scenario. The robustness of this method needs to be verified. In the future, we will obtain clinical data from different parts and scan parameters or simulate projections from different views for further experiments.

5. Conclusions

This study proposes a diffusion model for the artifact correction task of lung CBCT images. ScDiff achieves unsupervised learning by integrating the Match Module and the Diffusive Module into a cycle-consistency architecture while utilizing a large-step conditional diffusion process for efficient and accurate image sampling. In addition, the model introduces a SCConv module to handle redundant features, thereby improving network performance. Compared with existing artifact correction methods, the ScDiff model exhibits excellent performance and can generate high-fidelity and high-contrast images.

Author Contributions: Conceptualization, X.L. and Y.X.; methodology, Y.H.; software, Y.H.; validation, Y.H., X.L., and J.D.; formal analysis, Y.H.; investigation, G.D.; resources, G.D.; data curation, J.D.; writing—original draft preparation, Y.H.; writing—review and editing, G.D.; visualization, Y.H. and X.L.; supervision, Y.H.; project administration, Y.X.; funding acquisition, X.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by grants from the National Key Research and Develop Program of China (2023YFC2411502), National Natural Science Foundation of China (82202954, U20A20373, U21A20480).

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable

Data Availability Statement: https://www.cancerimagingarchive.net/.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Chen, H.; Zhang, Y.; Kalra, M.K.; Lin, F.; Chen, Y.; Liao, P.; Zhou, J.; Wang, G. Low-dose CT with a residual encoder-decoder convolutional neural network. *IEEE Trans. Med Imaging* **2017**, *36*, 2524–2535.
2. Bousse, A.; Kandarpa, V.S.S.; Rit, S.; Perelli, A.; Li, M.; Wang, G.; Zhou, J.; Wang, G. Systematic Review on Learning-based Spectral CT. *arXiv* **2023**, arXiv:2304.07588.
3. Nasrin, S.; Alom, M.Z.; Burada, R.; Taha, T.M.; Asari, V.K. Medical image denoising with recurrent residual u-net (r2u-net) base auto-encoder. In Proceedings of the 2019 IEEE National Aerospace and Electronics Conference (NAECON), Dayton, OH, USA, 15–19 July 2019 ; IEEE: Piscataway, NJ, USA, 2019; pp. 345–350.
4. Wu, W.; Hu, D.; Niu, C.; Yu, H.; Vardhanabhuti, V.; Wang, G. DRONE: Dual-domain residual-based optimization network for sparse-view CT reconstruction. *IEEE Trans. Med Imaging* **2021**, *40*, 3002–3014.
5. Guo, X.; Wu, W.; Duan, X.; Yu, H.; Chang, D.; He, P.; Wang, J.; Zhou, R.; Du, Y.; An, K. Nonlinear Deviation CorrecTion for Ring-Artifact Removal With Cone-Beam Computed Tomography. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 5019510.
6. Kida, S.; Nakamoto, T.; Nakano, M.; Nawa, K.; Haga, A.; Kotoku, J.; Yamashita, H.; Nakagawa, K. Cone beam computed tomography image quality improvement using a deep convolutional neural network. *Cureus* **2018**, *10*, e2548.
7. Chang.; Shaojie.; Chen.; Xi.; Duan.; Jiayu.; Mou.; Xuanqin. A CNN-based hybrid ring artifact reduction algorithm for CT images. *IEEE Trans. Radiat. Plasma Med Sci.* **2020**, *5*, 253–260.
8. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Commun. ACM* **2020**, *63*, 139–144.
9. Radford, A.; Metz, L.; Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv* **2015**, arXiv:1511.06434.
10. Xu, W.; Long, C.; Wang, R.; Wang, G. Drb-gan: A dynamic resblock generative adversarial network for artistic style transfer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual , 11–17 October 2021; pp. 6383–6392.
11. Huang, S.W.; Lin, C.T.; Chen, S.P.; Wu, Y.Y.; Hsu, P.H.; Lai, S.H. Auggan: Cross domain adaptation with gan-based data augmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 718–731.
12. Liang, X.; Chen, L.; Nguyen, D.; Zhou, Z.; Gu, X.; Yang, M.; Wang, J.; Jiang, S. Generating synthesized computed tomography (CT) from cone-beam computed tomography (CBCT) using CycleGAN for adaptive radiation therapy. *Phys. Med. Biol.* **2019**, *64*, 125002.
13. Dong, G.; Zhang, C.; Deng, L.; Zhu, Y.; Dai, J.; Song, L.; Meng, R.; Niu, T.; Liang, X.; Xie, Y. A deep unsupervised learning framework for the 4D CBCT artifact correction. *Phys. Med. Biol.* **2022**, *67*, 055012.
14. Wang, T.; Liu, X.; Dai, J.; Zhang, C.; He, W.; Liu, L.; Chan, Y.; He, Y.; Zhao, H.; Xie, Y.; et al. An unsupervised dual contrastive learning framework for scatter correction in cone-beam CT image. *Comput. Biol. Med.* **2023**, *165*, 107377.
15. Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; Chen, X. Improved techniques for training gans. *Adv. Neural Inf. Process. Syst.* **2016**, *29*.
16. Liu, X.; Xie, Y.; Diao, S.; Tan, S.; Liang, X. Unsupervised CT Metal Artifact Reduction by Plugging Diffusion Priors in Dual Domains. *IEEE Trans. Med. Imaging* **2024**, *43*, 3533–3545.
17. Özbey, M.; Dalmaz, O.; Dar, S.U.; Bedel, H.A.; Özturk, Ş; Güngör, A.; Çukur, T. Unsupervised medical image translation with adversarial diffusion models. *IEEE Trans. Med. Imaging* **2023**, *42*, 3524–3539.

18. Li, X.; Ren, Y.; Jin, X.; Lan, C.; Wang, X.; Zeng, W.; Wang, X.; Chen, Z. Diffusion Models for Image Restoration and Enhancement–A Comprehensive Survey. *arXiv* **2023**, arXiv:2308.09388.
19. Li, J.; Wen, Y.; He, L. SCConv: Spatial and Channel Reconstruction Convolution for Feature Redundancy. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 6153–6162.
20. Roman, N.O.; Shepherd, W.; Mukhopadhyay, N.; Hugo, G.D.; Weiss, E. Interfractional positional variability offiducial markers and primarytumors in locally advanced non-small-cell lung cancer during audiovisual biofeedback radiotherapy. *Int. J. Radiat. Oncol. Biol. Phys.* **2012**, *83*, 66–72.
21. Balik, S.; Weiss, E.; Jan N.; Roman, N.; Sleeman, W.C.; Fatyga, M.; Christensen, G.E.; Zhang, C.; Murphy, M.J.; Lu, J.; et al. Evaluation of 4-dimensional computed tomographyto 4-dimensional cone-beam computed tomographydeformable image registration for lung cancer adaptive radiation therapy. *Int. J. Radiat. Oncol. Biol. Phys.* **2013**, *86*, 2–9.
22. Clark, K.; Vendt, B.; Smith, K.; Freymann, J.; Kirby, J.; Koppel, P.; Moore, S.; Phillips, S.; Maffitt, D.; Pringle, M.; et al. The cancer imaging archive (TCIA): Maintaining and operating a public information repository. *J. Digit. Imaging* **2013**, *26*, 45–57.
23. Hugo, G.D.; Weiss, E.; Sleeman, W.C.; Balik, S.; Keall, P.J.; Lu, J.; Williamson, J.F. Data from 4D Lung Imaging of NSCLC Patients. *Cancer Imaging Arch.* **2016**. http://doi.org/10.7937/K9/TCIA.2016.ELN8YGLE.
24. Hugo, G.D.; Weiss, E.; Sleeman, W.C.; Balik, S.; Keall, P.J.; Lu, J.; Williamson, J.F. Alongitudinal four-dimensional computed tomography and cone beam computed tomography dataset for image-guided radiation therapy research in lung cancer. *Med. Phys.* **2017**, *44*, 62–71.
25. Bera, A.K.; Jarque, C.M.; Lee, L.F. Testing the normality assumption in limited dependent variable models. *Int. Econ. Rev.* **1984**, *25*, 563–578.
26. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27*.
27. Dar, S.U.; Yurt, M.; Karacan, L.; Erdem, A.; Erdem, E.; Cukur, T. Image synthesis in multi-contrast MRI with conditional generative adversarial networks. *IEEE Trans. Med Imaging* **2019**, *38*, 2375–2388.
28. Mescheder, L.; Geiger, A.; Nowozin, S. Which training methods for GANs do actually converge? In Proceedings of the International Conference on Machine Learning. PMLR, Stockholm, Sweden, 10–15 July 2018; pp. 3481–3490.
29. Zhang, W.; Yang, L.; Geng, S.; Hong, S. Self-supervised time series representation learning via cross reconstruction transformer. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, 35, 16129–16138.
30. Zhang, T.; Qi, G.-J.; Xiao, B.; Wang, J. Interleaved group convolutions for deep neural networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
31. Hua, B.S.; Tran, M.K.; Yeung, S.K. Pointwise convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 984–993.
32. Li, X.; Wang, W.; Hu, X.; Yang, J. Selective kernel networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 510–519.
33. Chai, T.; Draxler, R.R. Root mean square error (RMSE) or mean absolute error (MAE). *Geosci. Model Dev. Discuss.* **2014**, *7*, 1525–1534.
34. Poobathy, D.; Chezian, R.M. Edge detection operators: Peak signal to noise ratio based comparison. *IJ Image Graph. Signal Process.* **2014**, *10*, 55–61.
35. Chen, M.J.; Bovik, A.C. Fast structural similarity index algorithm. *J. Real-Time Image Process.* **2011**, *6*, 281–287.
36. Li, Y.; Shao, H.-C.; Liang, X.; Chen, L.; Li, R.; Jiang, S.; Wang, J.; Zhang, Y. Zero-shot Medical Image Translation via Frequency-Guided Diffusion Models. *arXiv* **2023**, arXiv:2304.02742.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.