DCT-Net: Dual-branch CT Reconstruction from Orthogonal X-rays with Diffusion Model and Contrastive Learning *

Zhiyu Zhang[1], Cong Shen[1✉], Jijun Tang[2], Zhijun Liao[3]

[1] Tianjin University of Technology, Tianjin, China
congshen@email.tjut.edu.cn
[2] Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, Guangdong, China
[3] Fujian Medical University, Fujian, Fuzhou, China

Abstract. Computed tomography (CT) reconstruction from X-ray images possesses significant advantages, including lower radiation exposure, reduced costs, and better accessibility than direct CT imaging. However, insufficient effective input samples caused by data volume under the moderate level or occlusion of partial soft tissues by skeletal structures in X-rays often hold back achieving high-quality image reconstruction. Additionally, contrasted with voxel-level differences, the texture and structure features are significant for image reconstruction. In virtue of these challenges, this study proposes an efficient approach named **D**ual-branch **CT Net**work (DCT-Net). It first integrates a conditional diffusion model for data augmentation, which mitigates data scarcity and achieves bone suppression. Subsequently, a dual-branch network in DCT-Net is leveraged to parallel process both augmented and raw data. In the framework, a perceptual loss based on high-level semantic features performs as the contrastive loss. Furthermore, it combines the voxel-level and adversarial losses to optimize the generator. However, the discriminator optimization only depends on the adversarial loss. Experimental results on two public datasets demonstrate that DCT-Net outperforms the state-of-the-art works, appearing to have promising potential among clinical applications.

Keywords: Computed tomography · X-rays · Dual-branch network · 3D reconstruction · Diffusion model · Contrastive learning.

1 Introduction

Computed tomography (CT) generates cross-sectional images from multi-angle X-ray projections, providing visualization of bones and soft tissues with higher resolution than X-ray images. However, considering its risky dose of radiation [1,2], low-dose scanning approaches could be viewed as the alternative manners [3–5], which still require hundreds of X-ray projections from a CT scanner.

* This work is sponsored in part by National Natural Science Foundation of China (No.62106175, 62020106004 and 92048301).

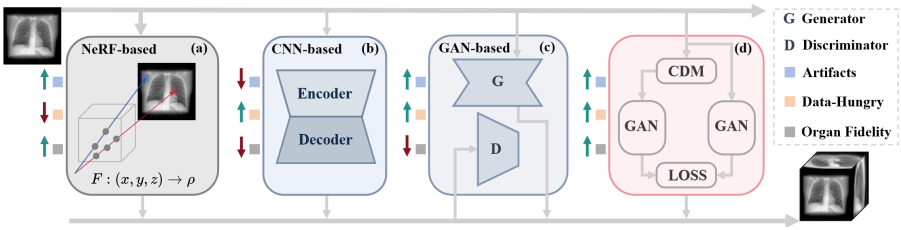

Fig. 1. Motivation for data augmentation in CT reconstruction. The arrows in different colors refer to the performance of the corresponding strategy. The red rounded rectangle refers to the data augmentation strategy in this study.

Recently, CT reconstruction from X-ray images has arisen, focusing on its effectiveness in cost-saving and low-dose imaging [6]. Generally, the encountered problems for CT reconstruction with X-rays could be categorized into the sparse view [7–11] and limited sample size. The prior usually adopts neural radiance fields (NeRF) to construct a continuous 3D volume representation. The latter generally employs a convolutional neural network (CNN) and generative adversarial network (GAN) to learn the mapping from 2D to 3D, as shown in Fig. 1. Neither (a) nor (b) and (c) could handle the requirement for a large amount of input data, the reconstruction of organ details, and artifacts.

For the CNN-based methods in Fig. 1(b), they extract image features through multiple convolutional layers [12], and learn the mapping from X-rays to CT [13–16]. However, the local receptive fields of convolutional operations also limit the capturing of global 3D structural information, leading to artifacts or anatomical structures missing. Contrastively, NeRF employs a continuous function [9, 17], which could produce relatively high-resolution results by leveraging X-rays taken from different angles [8, 18]. NeRF estimates density values based on the light propagation in space and the interactions among surrounding tissues. Nevertheless, NeRF-based methods are data-hungry and highly rely on the amount of multi-view X-rays.

GAN utilizes adversarial training, which adopts generator and discriminator to compete against each other to produce more realistic images [19–24]. Additionally, GAN could effectively capture high-dimensional data distributions, making it well-suited for CT generation [20]. However, it often focuses on voxel-level restoration but ignores high-level semantic information. Moreover, due to the occlusion of soft tissues by the skeletal structure, the reconstruction of organ details could be viewed as a challenge.

Inspired by GAN and contrastive learning, this study proposes a refined framework DCT-Net, which integrates a pre-trained conditional diffusion model (CDM) and contrastive loss as Fig. 1(d). Initially, the CDM based on the orthogonal X-rays is employed to generate bone-suppressed X-rays, which reflect the positions and features of the internal organs [25]. Then, a dual-branch framework is designed to play a role in measuring the effect of data augmentation. Subsequently, to leverage the high-level semantic information of structure and texture

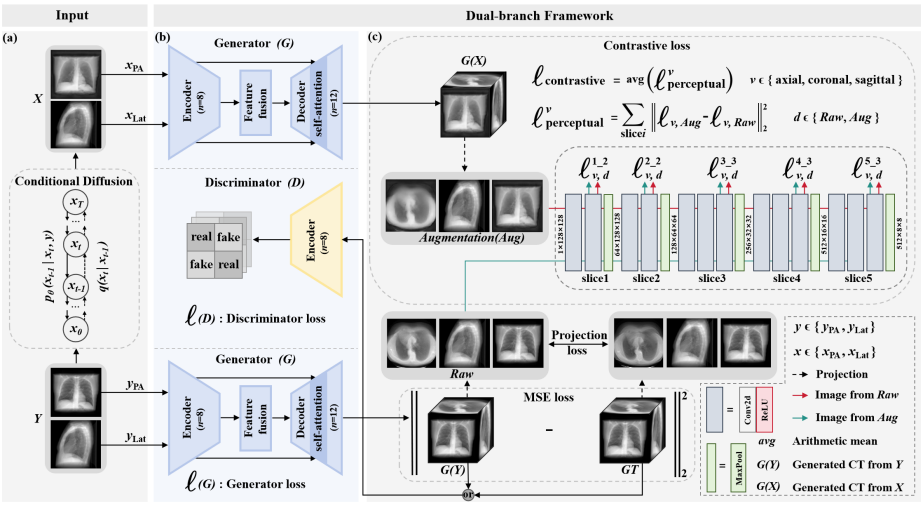

Fig. 2. The schematic of the DCT-Net framework. PA denotes the posterior-anterior view; Lat denotes the lateral view. (a) Generate bone-suppressed X-ray images from a pre-trained conditional diffusion model conditioned on input X-ray images. (b) Dual-generator reconstruction GAN processes the X-ray images. (c) Multi-path loss module.

embedded in the CT generated by the proposed dual-branch network, perceptual loss is applied as contrastive loss [26]. In addition, the final multi-path loss module also integrates reconstruction loss, projection loss, and adversarial loss to enhance the robustness and generalization ability. The main contributions of this work are listed below:

1. A conditional diffusion model is used in generating bone-suppressed X-rays to improve the visibility of the internal organ structures in CT and overcome data scarcity of input X-rays;
2. A dual-branch framework for CT reconstruction is proposed, which separately processes the original and bone-suppressed X-rays via a dual-generator reconstruction GAN, improving reconstruction accuracy and robustness;
3. A multi-path loss module is applied to balance the voxel-level details and high-level semantic features.

2 Method

The framework proposed in this study is named **D**ual-branch **CT Net**work (DCT-Net). As illustrated in Fig. 2, it leverages a pair of input orthogonal X-ray images $\langle y_{\mathrm{PA}}, y_{\mathrm{Lat}} \rangle$ to generate CT images.

2.1 Bone Suppression Conditional Diffusion Model

Due to the data scarcity of input X-rays and the occlusion of internal organs caused by bones, a CDM is employed for data augmentation, enabling the model to focus on the texture, structure, and location of internal organs. Given that x and y are the images in the image set X and Y, respectively. Then, DCT-Net generates bone-suppressed X-rays $X = \{x_{\mathrm{PA}}, x_{\mathrm{Lat}}\}$ from orthogonal X-rays $Y = \{y_{\mathrm{PA}}, y_{\mathrm{Lat}}\}$.

Diffusion Model. Diffusion models include forward and reverse processes [25, 27, 28]. The forward process q gradually perturbs an initial data distribution $x_0 \sim q(x_0)$ using a predefined Markov chain, progressively adding Gaussian noise over T steps until $x_T \sim \mathcal{N}(0, \mathbf{I})$, which shown as (1):

$$q(x_t \mid x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t \mathbf{I}\right) \tag{1}$$

where β_t denotes the noise scheduling coefficient at timestep t, controlling the noise injection rate; $\mathbf{I}$ denotes the identity matrix. The reverse process p_θ aims to recover the original data x_0 from the noise x_T, which is formulated as (2):

$$p_\theta\left(x_{t-1} \mid x_t\right) = \mathcal{N}\left(x_{t-1}; \mu_\theta\left(x_t, t\right), \Sigma_\theta\left(x_t, t\right)\right) \tag{2}$$

where μ_θ and Σ_θ denote the mean and covariance derived from a U-Net-based noise prediction network ϵ_θ parameterized by θ [29].

Conditional Diffusion Model. As the bone-suppressed X-ray x is generated according to the corresponding X-ray y, the reverse process is formulated as $p_\theta(x_{0:T} \mid y)$, with the original forward process q preserved. The conditional reverse transition could be (3):

$$p_\theta(x_{t-1} \mid x_t, y) = \mathcal{N}\left(x_{t-1}; \mu_\theta(x_t, y, t), \Sigma_\theta(x_t, y, t)\right) \tag{3}$$

During the training process, pairs $(x_0, y) \sim q(x_0, y)$ are sampled, and ϵ_θ learns to predict noise conditioned on y, where x_0 denotes the ground truth bone-suppressed X-ray. The joint distribution could be formulated as (4):

$$p_\theta(x_{0:T} \mid y) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1} \mid x_t, y) \tag{4}$$

2.2 Dual-generator Reconstruction GAN

The GAN in DCT-Net uses two generators to process X and Y, of which the discriminator optimizes one of the generators that processes Y, shown as Fig. 2(b). Specifically, $\mathbf{p}_{\mathrm{LA}} \in \{x_{\mathrm{LA}}, y_{\mathrm{LA}}\}$ and $\mathbf{p}_{\mathrm{Lat}} \in \{x_{\mathrm{Lat}}, y_{\mathrm{Lat}}\}$. $G(X)$ and $G(Y)$ denote the generated CT from X and Y, respectively.

Generator. To address the limitation of capturing the latent relationship between two views with two encoders that increases the complexity of model [21–23], each generator in DCT-Net is based on an encoder-decoder, involving a feature fusion module γ and self-attention. Ying et al. proposed a skip connection

to adaptively transmit the low-level features during the encoding process [22]. To this end, the fused features of orthogonal X-rays could be defined as (5):

$$\mathbf{f}_{fusion} = \gamma(\mathbf{f}_1, \mathbf{f}_2) = \mathrm{FC}([\mathrm{EC}(\mathbf{p}_{\mathrm{LA}}), \mathrm{EC}(\mathbf{p}_{\mathrm{Lat}})]) \tag{5}$$

where $\mathrm{FC}(\cdot)$ and $\mathrm{EC}(\cdot)$ denote fully connected layer and encoder, respectively. $[\mathrm{EC}(\mathbf{p}_{\mathrm{LA}}), \mathrm{EC}(\mathbf{p}_{\mathrm{Lat}})]$ represents the concatenation of the features from $\mathbf{p}_{\mathrm{LA}}$ and $\mathbf{p}_{\mathrm{Lat}}$. For the decoder, self-attention is incorporated with each layer to enhance the capture of long-range dependencies in the reconstructed CT features, as shown in (6):

$$\mathrm{DC}_l = \mathrm{SA}\left(\mathrm{Deconv}\left(\mathrm{DC}_{l-1}\right)\right), \quad l = 1, 2, \ldots, L \tag{6}$$

where $\mathrm{Deconv}(\cdot)$ denotes deconvolution; $\mathrm{SA}(\cdot)$ represents self-attention; and DC_l denotes the output of the l-th decoding layer. The generated CT denotes as DC_L.

Discriminator. The discriminator in GAN outputs a scalar to determine the fidelity of an input sample [30]. This setting provides a unified judgment for the image without considering the significant bias on average for local areas. Hence, the output of the discriminator is constructed as (7), allowing the generator to obtain fine-grained feedback:

$$D(\cdot|Y) = \{d_{j,v}\}_{m \times n} \tag{7}$$

where '$\cdot$' denotes $G(Y)$ or the CT ground truth GT; $v \in \{axial, coronal, sagittal\}$; D represents the discriminator; m and n denote the patches of one CT horizontal and vertical slice, respectively; $d_{j,v}$ represents the probability of sub-region in j-th slice of $G(Y)$ from the sample distribution, as viewed from v.

2.3 Multi-path Loss Module

In DCT-Net, a multi-path loss module is designed to integrate perceptual loss, voxel-level loss, and adversarial loss.

Contrastive loss. Regarding $G(X)$ is expected to suppress skeletal information, rather than enforcing a strict pixel-wise match with $G(Y)$, DCT-Net introduces the perceptual loss as the contrastive loss to measure high-level semantic and perceptual differences between them.

Specifically, $G(Y)$ and $G(X)$ are projected onto standard anatomical planes to obtain the *Raw* and *Aug* images, respectively. Suppose that $\phi_i^v(z)$ is the ReLU activation of the i-th layer in a network ϕ for an image z under view v. When i corresponds to a convolutional layer, $\phi_i^v(z)$ is a feature map of shape $C_i \times H_i \times W_i$. The perceptual loss between *Raw* and *Aug* on the same plane is calculated as (8):

$$\ell^v(Raw, Aug) = \sum_{i=1}^{N_{\mathrm{s}}} \frac{1}{C_i H_i W_i} \left\| \phi_i^v(Raw) - \phi_i^v(Aug) \right\|_2^2 \tag{8}$$

where N_S denotes the number of slices in network ϕ. $\phi_i^v(Raw)$ and $\phi_i^v(Aug)$ represent the activation values in i-th layer of ϕ, extracted from *Raw* and *Aug*

on the projection plane with the same view v. Contrastive loss ℓ_{contr} is computed as the arithmetic mean of the perceptual loss across the three planes.
Voxel-level Loss. It includes reconstruction loss ℓ_{recon} and projection loss ℓ_{proj}, ensuring voxel-wise consistency not only in 3D space but also across projected 2D slices. Reconstruction loss employs mean squared error (MSE) to minimize voxel-wise differences. Projection loss enforces consistency by computing the average L1 loss across axial, coronal, and sagittal planes.
Adversarial Loss. Following the LSGAN [31], the adversarial loss of DCT-Net is defied as (9):

$$\begin{aligned}\ell(D) = \frac{1}{2}\Big[&\mathbb{E}_{GT\sim p(\mathrm{CT})}(D(GT \mid Y)-1)^2 \\ &+ \mathbb{E}_{Y\sim p(\mathrm{Xray})}(D(G(Y) \mid Y)-0)^2\Big]\end{aligned} \tag{9a}$$

$$\ell(G) = \frac{1}{2}\mathbb{E}_{Y\sim p(\mathrm{Xray})}\left[(D(G(Y) \mid Y)-1)^2\right] \tag{9b}$$

where $GT \sim p(\mathrm{CT})$ represents the real CT samples, and $Y \sim p(\mathrm{Xray})$ represents the input X-rays samples. Discriminator loss $\ell(D)$ comprises two terms to encourage it to classify GT and penalize it for classifying $G(Y)$. Generator loss function $\ell(G)$ encourages the discriminator to classify $G(Y)$ with high fidelity.
Overall Training Objective. It is defined as (10):

$$D^* = \arg\min_D \lambda_1 \ell(D) \tag{10a}$$

$$G^* = \arg\min_G \left[\lambda_1 \ell(G) + \lambda_2 \ell_{contr} + \lambda_3 \ell_{recon} + \lambda_4 \ell_{proj}\right] \tag{10b}$$

where λ_1, λ_2, λ_3, and λ_4 are hyper-parameters to adjust the weight of loss terms.

3 Experiments

3.1 Datasets and Experimental Settings

LIDC-IDRI Dataset. The dataset contains 1,018 lung CT samples [32]. In this study, 854 samples are used for training, 72 for validation, and 92 for testing.
CTspine1K Dataset. 784 CT images [33] are selected from this dataset and divided into 650 training samples, 60 validation samples, and 74 testing samples.
Implementations and Training Settings. In the experiments, the input X-rays synthesized by the digitally reconstructed radiographs technology are resized to 128×128 [34], and the output CT is resized to 128×128×128. The implementation is based on the PyTorch framework with Adam optimizer, where the initial learning rate is set to 0.0002 and momentum parameters $\beta_1 = 0.5$, $\beta_2 = 0.999$. Furthermore, due to the constraints of GPU memory, the batch size is set as 1. The training process is carried out for 100 epochs to ensure stable convergence based on the validation set. All the experiments are conducted on a workstation with an NVIDIA GeForce RTX 3090 Ti 24GB GPU card.

Table 1. Experimental results and ablation study on LIDC-IDRI dataset.

Method	**PSNR** (↑)	**SSIM** (↑)	**LPIPS** (↓)
PerX2CT [37]	27.248 ± 0.004	0.637 ± 0.005	0.253 ± 0.003
X2CT [22]	26.769 ± 0.040	0.622 ± 0.001	0.307 ± 0.002
RT-SRTS [38]	25.311 ± 0.010	0.570 ± 0.001	0.323 ± 0.003
X-Recon [23]	27.120 ± 0.001	0.619 ± 0.002	0.292 ± 0.004
SdCT-GAN [21]	27.054 ± 0.010	0.596 ± 0.003	0.322 ± 0.020
w/o Dual-branch	27.226 ± 0.030	0.620 ± 0.005	0.285 ± 0.001
w/o Recon-GAN	27.562 ± 0.001	0.648 ± 0.008	0.251 ± 0.004
w/o $\ell_{perceptual}$	27.602 ± 0.003	0.651 ± 0.003	0.283 ± 0.001
DCT-Net (ours)	$\mathbf{27.928 \pm 0.040}$	$\mathbf{0.703 \pm 0.001}$	$\mathbf{0.221 \pm 0.004}$

Table 2. Experimental results on CTspine1K dataset.

Method	**PSNR** (↑)	**SSIM** (↑)	**LPIPS** (↓)	**Params** (M)
PerX2CT [37]	25.575 ± 0.010	0.614 ± 0.001	0.282 ± 0.003	71.59
X2CT [22]	24.927 ± 0.001	0.616 ± 0.008	0.334 ± 0.001	73.15
RT-SRTS [38]	23.518 ± 0.020	0.523 ± 0.002	0.348 ± 0.005	546.26
X-Recon [23]	24.715 ± 0.010	0.594 ± 0.003	0.330 ± 0.001	101.74
SdCT-GAN [21]	25.124 ± 0.040	0.541 ± 0.003	0.336 ± 0.002	77.32
DCT-Net (ours)	$\mathbf{26.600 \pm 0.003}$	$\mathbf{0.691 \pm 0.010}$	$\mathbf{0.234 \pm 0.030}$	**60.70**

Evaluation Metric. Peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) are adopted in this work [35]. Moreover, learned perceptual image patch similarity (LPIPS) is introduced to focus on latent semantic perception [36].

3.2 Comparison with State-of-the-art Methods

Table 1 presents the comparison results on the LIDC-IDRI dataset. DCT-Net outperforms other methods in PSNR, SSIM, and LPIPS. Compared with the current SOTA method PerX2CT, DCT-Net achieves improvements of +0.68 in PSNR, +0.066 in SSIM, and −0.032 in LPIPS. In particular, PerX2CT adopts slice-wise processing in CT reconstruction [37], ignoring cross-slice continuity required in 3D structural consistency. To evaluate the stability and generalization of DCT-Net and baseline methods, experiments on the CTspine1K pose more reconstruction challenges due to the lesions in the dataset. As shown in Table 2, DCT-Net achieves the best performance while maintaining the most stable metrics with merely -1.328 PSNR and +0.013 LPIPS variation. Furthermore, DCT-Net requires the fewest parameters with only 60.7M, demonstrating higher computational efficiency.

Visualization results are shown in Fig. 3 where DCT-Net demonstrates superior reconstruction performance, particularly in preserving anatomical details such as lung structures, cardiac contours, and specific tracheal regions. This outcome is consistent with the results presented in Tables 1 and 2.

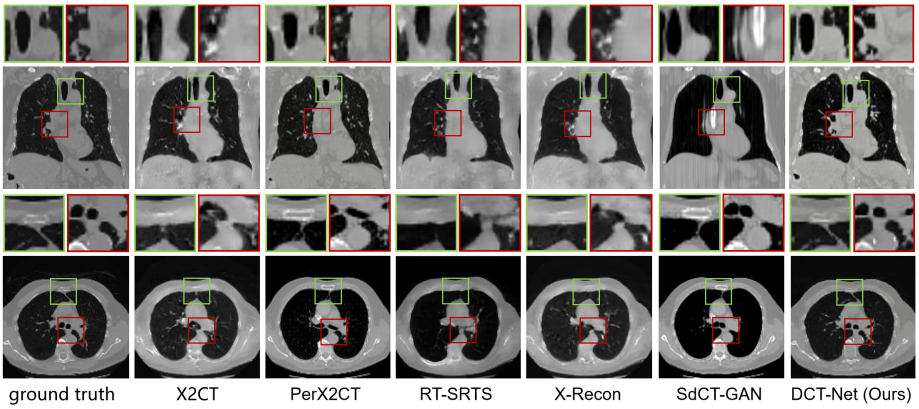

Fig. 3. Visual comparison of CT slices. The second and the last rows are the global CT slices, while the first and the third rows are the local details of global CT slices.

3.3 Ablation Study

To verify the effectiveness of dual-branch network, dual-generator reconstruction GAN (Recon-GAN), and perceptual loss in DCT-Net, the ablation results are presented in Table 1.

Firstly, the bone suppression CDM failed without the dual-branch network component. While PSNR and SSIM decrease by 0.702 and 0.083, LPIPS increases by 0.064 compared with the combination. It indicates that bone-suppressed X-rays could conduct data cleaning for input data, reducing the interference of bone artifacts. Meanwhile, the dual-branch network could leverage the complementary information from the output CTs of the Recon-GAN. Secondly, Recon-GAN is replaced with the GAN proposed by Ying et al. [22]. Results show that it declines across all three evaluation metrics, verifying the significance of DCT-Net improvements to the generator and discriminator. Lastly, perceptual loss replaced with projection loss is conducted. LPIPS increases by 0.062 in Table 1, indicating that perceptual loss tends to capture high-level semantic information while avoiding optimizing only with voxel-level differences.

4 Conclusion

In this study, a network named DCT-Net is proposed to achieve CT reconstruction from orthogonal X-rays. A conditional diffusion model is firstly employed to generate bone-suppressed X-rays, overcoming data scarcity and the occlusion of internal organs by bones. Then, the augmented and original X-rays are fed into a dual-generator reconstruction GAN, which incorporates a feature fusion module and self-attention. Next, the perceptual loss is selected as the contrastive loss in DCT-Net, focusing on restoring information about internal organs. Experimental results on the LIDC-IDRI and CTspine1K datasets demonstrate that DCT-Net

significantly outperforms existing benchmark methods. Future work will focus on data augmentation using self-supervised and semi-supervised methods for cases lacking samples.

Acknowledgements. The authors gratefully acknowledge the reviewers for their valuable comments and suggestions. We would also like to thank our colleagues.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Power, S., Moloney, F., Twomey, M., James, K., O'Connor, O., Maher, M.: Computed tomography and patient risk: facts, perceptions and uncertainties. World J. Radiol. **8**(12), 902–915 (2016)
2. Schmidt, C.: CT scans: balancing health risks and medical benefits. Environ. Health Perspect. **120**, A118–A121 (2012)
3. Zhang, Y., Cao, R., Xu, F., Zhang, R., Jiang, F., Meng, J., Ma, F., Guo, Y., Liu, J.: Side information-assisted low-dose CT reconstruction. IEEE Trans. Comput. Imaging **10**, 1080–1093 (2024)
4. Zhang, W., Huang, B., Chen, S., Xu, X., Wu, W., Liu, Q.: Low-rank angular prior guided multi-diffusion model for few-shot low-dose CT reconstruction. IEEE Trans. Comput. Imaging **10**, 1763–1774 (2024)
5. Jiang, J., Feng, Y., Xu, H., Zheng, J.: Low-dose CT reconstruction via optimization-inspired GAN. In: ICASSP. pp. 1–5 (2023)
6. Garg, Y., Seetharam, K., Sharma, M., Rohita, D.K., Nabi, W.: Role of deep learning in computed tomography. Cureus **15**(5), e39160 (2023)
7. Ding, C., Zhang, Q., Wang, G., Ye, X., Chen, Y.: Learned alternating minimization algorithm for dual-domain sparse-view CT reconstruction. In: MICCAI. vol. 14229, pp. 173–183. Springer (2023)
8. Cai, Y., Wang, J., Yuille, A., Zhou, Z., Wang, A.: Structure-aware sparse-view X-ray 3D reconstruction. In: CVPR. pp. 11174–11183 (2024)
9. Zha, R., Zhang, Y., Li, H.: NAF: neural attenuation fields for sparse-view CBCT reconstruction. In: MICCAI. vol. 13436, pp. 442–452. Springer (2022)
10. Cheng, Y., Li, Q., Li, R., Wang, T., Zhao, J., Yan, Q., Rehman, Z.U., Wang, L., Geng, Y.: LIR-Net: learnable iterative reconstruction network for fan beam CT sparse-view reconstruction. IEEE Trans. Comput. Imaging **10**, 181–195 (2024)
11. Ayad, I., Larue, N., Nguyen, M.K.: QN-Mixer: a quasi-newton MLP-mixer model for sparse-view CT reconstruction. In: CVPR. pp. 25317–25326 (2024)
12. Shen, C., Wang, X., Tang, J., Liao, Z.: DETA-Net: A dual encoder network with text-guided attention mechanism for skin-lesions segmentation. In: ICIC. pp. 28–40 (2023)
13. Chen, H., Zhang, Y., Zhang, W., Liao, P., Li, K., Zhou, J., Wang, G.: Low-dose CT via convolutional neural network. Biomed. Opt. Express. **8**(2), 679–694 (2017)
14. Jin, K., McCann, M., Froustey, E., Unser, M.: Deep CNN for inverse problems in imaging. IEEE Trans. Image Process. **26**(9), 4509–4522 (2017)

15. Dai, J., Dong, G., Zhang, C., He, W., Liu, L., Wang, T., Jiang, Y., Zhao, W., Zhao, X., Xie, Y., Liang, X.: Volumetric tumor tracking from a single cone-beam X-ray projection image enabled by deep learning. Med Image Anal. **91**, 102998 (2024)
16. Shao, H., Wang, J., Bai, T., Chun, J., Park, J.C., Jiang, S., Zhang, Y.: Real-time liver tumor localization via a single X-ray projection using deep graph neural network-assisted biomechanical modeling. Phys Med Biol. **67**(11) (2022)
17. Rückert, D., Wang, Y., Li, R., Idoughi, R., Heidrich, W.: NeAT: neural adaptive tomography. ACM Trans. Graph. **41**(4) (2022)
18. Corona-Figueroa, A., Frawley, J., Taylor, S., Bethapudi, S., Shum, H., Willcocks, C.: MedNeRF: medical neural radiance fields for reconstructing 3D-aware CT-projections from a single X-ray. In: EMBC. pp. 3843–3848 (2022)
19. Yang, Q., Yan, P., Zhang, Y., Yu, H., Shi, Y., Mou, X., Kalra, M., Zhang, Y., Sun, L., Wang, G.: Low-dose CT image denoising using GAN with Wasserstein distance and perceptual loss. IEEE Trans. Med. Imaging **37**(6), 1348–1357 (2018)
20. Cheng, S., Hong, Y., Chen, Q., Guo, J., Li, M., Zhang, Q.: CLCT-GAN: strong-weak contrastive learning for reconstructing CT images from radiographs. In: IJCNN. pp. 1–8 (2024)
21. Cheng, S., Chen, Q., Zhang, Q., Li, M., Alike, Y., Su, K., Wen, P.: SdCT-GAN: reconstructing CT from biplanar X-rays with self-driven GAN. arXiv preprint arXiv:2309.04960 (2023)
22. Ying, X., Guo, H., Ma, K., Wu, J., Weng, Z., Zheng, Y.: X2CT-GAN: reconstructing CT from biplanar X-rays with GAN. In: CVPR. pp. 10611–10620 (2019)
23. Wang, Y., Wang, K., Zhuo, Y., Shi, W., Shan, F., Liu, L.: X-Recon: learning-based patient-specific high-resolution CT reconstruction from orthogonal X-ray images. arXiv preprint arXiv:2407.15356 (2024)
24. Zeng, P., Zeng, X., Wang, Y., Zhou, L., Zu, C., Wu, X., Zhou, J., Shen, D.: Multi-modal long-short distance attention-based transformer-GAN for PET reconstruction with auxiliary MRI. IEEE Trans. Circuits Syst. Video Technol. pp. 1–14 (2025). https://doi.org/10.1109/TCSVT.2025.3545911
25. Chen, Z., Sun, Y., Ge, R., Qin, W., Pan, C., Deng, W., Liu, Z., Min, W., Elazab, A., Wan, X., Wang, C.: BS-Diff: effective bone suppression using conditional diffusion models from chest X-ray images. In: ISBI. pp. 1–5 (2024)
26. Justin, J., Alexandre, A., Li, F.F.: Perceptual losses for real-time style transfer and super-resolution. In: ECCV. vol. 9906. Springer, Cham (2016)
27. Friedrich, P., Wolleb, J., Bieder, F., Durrer, A., Cattin, P.C.: WDM: 3D wavelet diffusion models for high-resolution medical image synthesis. In: MICCAI. vol. 15224, pp. 11–21. Springer (2025)
28. Cui, J., Zeng, X., Zeng, P., Liu, B., Wu, X., Zhou, J., Wang, Y.: MCAD: Multi-modal Conditioned Adversarial Diffusion Model for High-Quality PET Image Reconstruction . In: MICCAI 2024. vol. LNCS 15007 (2024)
29. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: MICCAI. vol. 9351. Springer (2015)
30. Liu, W., Ding, H.: Solving low-dose CT reconstruction via GAN with local coherence. In: MICCAI. vol. 14229. Springer, Cham (2023)
31. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Smolley, S.P.: Least squares generative adversarial networks. In: ICCV. pp. 2813–2821 (2017)
32. Armato, S.G.I., McLennan, G., Bidaut, L., et al.: The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on ct scans. Med. Phys. **38**(2), 915–931 (2011)

33. Deng, Y., Wang, C., Hui, Y., et al.: CTSpine1K: a large-scale dataset for spinal vertebrae segmentation in computed tomography. arXiv preprint arXiv:2105.14711 (2021)
34. Milickovic, N., Baltast, D., Giannouli, S., Lahanas, M., Zamboglou, N.: CT imaging based digitally reconstructed radiographs and their application in brachytherapy. Phys. Med. Biol. **45**(10), 2787–2800 (2000)
35. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
36. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR. pp. 586–595 (2018)
37. Kyung, D., Jo, K., Choo, J., Lee, J., Choi, E.: Perspective projection-based 3D CT reconstruction from biplanar X-rays. In: ICASSP. pp. 1–5. IEEE (2023)
38. Zhu, M., Fu, Q., Liu, B., Zhang, M., Li, B., Luo, X., Zhou, F.: RT-SRTS: angle-agnostic real-time simultaneous 3D reconstruction and tumor segmentation from single X-ray projection. Comput. Biol. Med. **173**, 108390 (2024)