ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10096296

PERSPECTIVE PROJECTION-BASED 3D CT RECONSTRUCTION FROM BIPLANAR X-RAYS

Daeun Kyung[1,*] *Kyungmin Jo*[1,*] *Jaegul Choo*[1] *Joonseok Lee*[2,3] *Edward Choi*[1,†]

[1]KAIST [2]Seoul National University [3]Google Research

ABSTRACT

X-ray computed tomography (CT) is one of the most common imaging techniques used to diagnose various diseases in the medical field. Its high contrast sensitivity and spatial resolution allow the physician to observe details of body parts such as bones, soft tissue, blood vessels, etc. As it involves potentially harmful radiation exposure to patients and surgeons, however, reconstructing 3D CT volume from perpendicular 2D X-ray images is considered a promising alternative, thanks to its lower radiation risk and better accessibility. This is highly challenging though, since it requires reconstruction of 3D anatomical information from 2D images with limited views, where all the information is overlapped. In this paper, we propose PerX2CT, a novel CT reconstruction framework from X-ray that reflects the perspective projection scheme. Our proposed method provides a different combination of features for each coordinate which implicitly allows the model to obtain information about the 3D location. We reveal the potential to reconstruct the selected part of CT with high resolution by properly using the coordinate-wise local and global features. Our approach shows potential for use in clinical applications with low computational complexity and fast inference time, demonstrating superior performance than baselines in multiple evaluation metrics.

Index Terms— X-ray computed tomography, CT reconstruction

1. INTRODUCTION

X-ray computed tomography (CT) is a medical imaging technique that produces cross-sectional images of the body from multi-view X-ray projection data scanned around the patient. Due to its advantages of having high spatial and density resolution, it is widely used in the medical domain. Specifically, CT is helpful to diagnose diseases as it simultaneously shows details of body parts such as bones, soft tissue. However, CT scans have the disadvantage of incurring more radiation exposure than other medical imaging techniques [1, 2, 3].

Previous work made effort to reduce the radiation dose by reducing the number of CT projections while maintaining the high quality of reconstruction based on traditional methods [4] or deep learning approaches [5, 6, 7, 8]. However, these studies are limited to only marginally reducing the radiation dose, still requiring hundreds of projection images obtained from CT scanners to ensure high quality of images. In addition, they still need a CT scanner which is less accessible than X-ray machines. Efforts to resolve these problems have led to attempts to reconstruct CT using X-rays obtained from traditional X-ray machines [9, 10, 11, 12, 13, 14, 15, 16].

X-ray imaging is the most frequently performed radiographic examination in hospitals, as it needs less physical and economic burden on the patient. In particular, X-rays have less radiation exposure than CT and do not require additional preparation such as contrast agents. While CT requires the patient to lie on a cylindrical device, X-rays do not have these limitations, making them easier to use for patients with reduced mobility. Thus, it will be desirable to reconstruct 3D internal body information from X-rays.

The primary challenge of reconstructing the CT volume from the X-ray is the lack of depth information; estimating 3D structures from 2D data is a well-known ill-posed problem [17]. The 3D reconstruction problem gets more ambiguous as we use fewer X-rays, making it harder to solve with traditional CT reconstruction approaches. Recently, large-scale data and deep learning models [9, 10, 13, 14, 16] powered to tackle this ultra-sparse view reconstruction problem by learning prior knowledge of human anatomy.

Some studies use a 2D-to-3D structure consisting of a 2D encoder and a 3D decoder to provide depth information from a few X-ray images [10, 13, 14, 16]. These models commonly replicate the 2D feature map on the depth axis because they assume that X-rays are generated by orthogonal projection (Figure 2). However, since X-rays are generated by the perspective projection, the orthogonal projection places the features of X-rays in the wrong position. To provide accurate features to the model to reconstruct CT, we place the 2D X-ray feature in a 3D space using the perspective projection method, which is how X-ray images are generated. Recent studies [10, 13, 14] utilize segmentation maps to improve reconstruction quality. A significant drawback of these approaches is the necessity of a segmentation map, which is laborious to obtain. In contrast, our model reconstructs the 3D CT volume using only two X-rays (PA and lateral view images) without additional information. We utilize two X-rays as input since a single X-ray image has insufficient depth information for accurate reconstruction of a 3D volume [15, 16].

Another limitation of previous studies is that they require high computational cost due to the heavy 3D decoder. Since the orthogonal projection places the same features across the entire depth axis, the model receives deficient information. To compensate this shortage, the previous models reconstruct CT volume at the same time using the 3D decoder, incurring significant computational cost. However, we show that the 2D decoder is sufficient to reconstruct accurate CTs than the 3D decoders with the perspective projection, better suited for the X-ray image. With a 2D decoder, the time complexity will be significantly reduced and each CT slice can be individually reconstructed efficiently. Nonetheless, the 2D decoder may lack information on the whole body compared to the 3D decoder. We thus additionally provide global features of the X-rays and the position of each voxel.

In this work, we propose a simple yet efficient CT reconstruc-

[*] The first two authors contributed equally.

[†] Corresponding author

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant (No.2019-0-00075, No.2021-0-02068), Korea Health Industry Development Institute (KHIDI) grant (No.HI21C1138), Korea Medical Device Development Fund grant (Project Number: 1711138160, KMDF_PR_20200901_0097), and National Research Foundation of Korea (NRF) grant (NRF-2021H1D3A2A03038607) funded by the Korea government(MSIT, MOTIE, MOHW, MFDS).

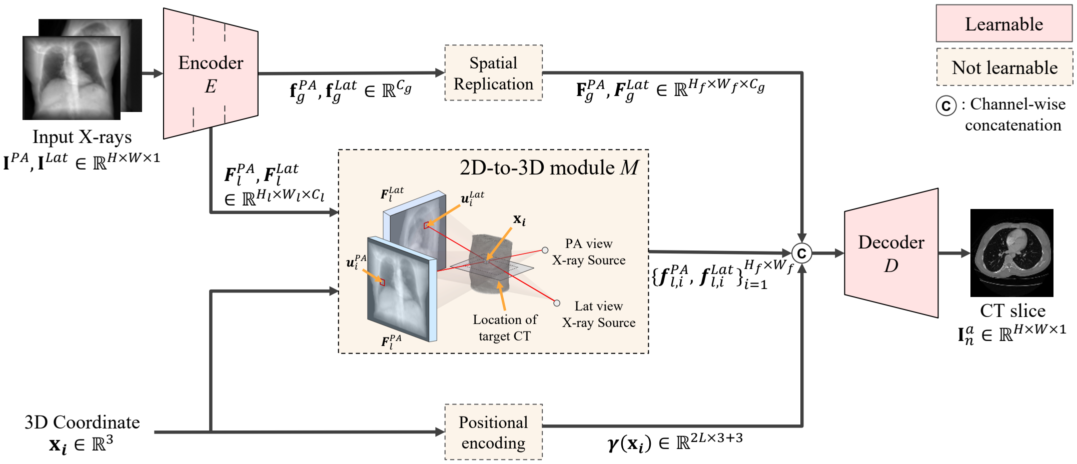
Learnable
Not learnable
Ⓒ : Channel-wise concatenation
Input X-rays
$\mathbf{I}^{PA}, \mathbf{I}^{Lat} \in \mathbb{R}^{H\times W\times 1}$
Encoder
E
$\mathbf{f}_g^{PA}, \mathbf{f}_g^{Lat} \in \mathbb{R}^{C_g}$
Spatial Replication
$\mathbf{F}_g^{PA}, \mathbf{F}_g^{Lat} \in \mathbb{R}^{H_f\times W_f\times C_g}$
$\mathbf{F}_l^{PA}, \mathbf{F}_l^{Lat}$
$\in \mathbb{R}^{H_l\times W_l\times C_l}$
2D-to-3D module M
$\mathbf{F}_l^{Lat}$
$\mathbf{u}_i^{Lat}$
$\mathbf{x}_i$
$\mathbf{u}_i^{PA}$
PA view X-ray Source
Lat view X-ray Source
Location of target CT
$\mathbf{F}_l^{PA}$
$\{\mathbf{f}_{l,i}^{PA}, \mathbf{f}_{l,i}^{Lat}\}_{i=1}^{H_f\times W_f}$
Decoder
D
CT slice
$\mathbf{I}_n^a \in \mathbb{R}^{H\times W\times 1}$
3D Coordinate
$\mathbf{x}_i \in \mathbb{R}^3$
Positional encoding
$\gamma(\mathbf{x}_i) \in \mathbb{R}^{2L\times 3+3}$

Figure 1: Overview of PerX2CT. The encoder extracts the local and global features of two X-rays and *2D-to-3D module* extracts the local latent vector of target slice. The decoder reconstructs the output patch using the corresponding local, global feature map and position encoding.

tion framework that reflects the X-ray perspective projection scheme, named PerX2CT. The proposed method significantly improves the visual quality of reconstructed CT by accurately placing the information acquired from 2D X-rays in the 3D space. With coordinate-wise features in the 3D space and the 2D decoder, we achieve not only a reduced model complexity but also a 10x faster inference time. In addition, it can perform partial reconstruction for arbitrary selected patches with high resolution thanks to its ability to utilize 3D positional information.

2. METHODOLOGY

In this part, we present our proposed CT reconstruction framework from X-rays taken from two perpendicular views, PA and lateral (Figure 1). We extract image features from X-rays and place them through the perspective projection method. Then, we reconstruct the target CT slice properly utilizing both local and global features.

2.1. Feature Extraction from X-rays

We extract image features from two X-rays with an encoder E_θ and place them in a 3D space to generate feature maps according to the target CT slice. To maximize the richness of information while minimizing the number of input X-ray images, we input two images of perpendicular views $\mathbf{I}^{PA}, \mathbf{I}^{Lat} \in \mathbb{R}^{H\times W\times 1}$ to E_θ, where $H \times W$ denote the spatial resolution of X-ray images. For simplicity, we denote the view of an image as $v \in \{PA, Lat\}$.

We encode local and global features for each view to represent the voxel-specified information and whole-body information, respectively. Since each pixel value of the X-ray is independently calculated using voxels through which each X-ray beam passes, appropriate allocation of local feature using the path of each ray helps reconstructing each voxel of CT. However, since local features only provide position-specified information, additional global features of X-rays help the model understand the whole body of a patient. Thus, we extract low-level local feature maps $\mathbf{F}_l^v \in \mathbb{R}^{H_l\times W_l\times C_l}$ and high-level global feature vectors $\mathbf{f}_g^v \in \mathbb{R}^{C_g}$ from the encoder and provide them to the decoder as input. C_l and C_g are the channel size of each feature and $(H_l \times W_l)$ denotes the spatial dimension of the local feature map.

$$E_\theta(\mathbf{I}^v) = (\mathbf{F}_l^v, \mathbf{f}_g^v), \tag{1}$$

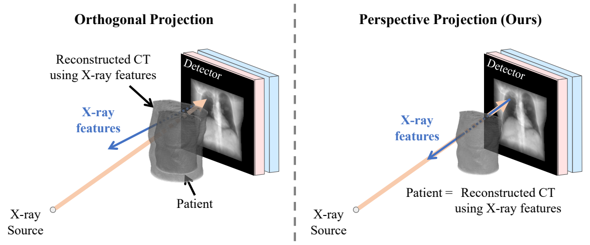
Orthogonal Projection
Reconstructed CT using X-ray features
Detector
X-ray features
Patient
X-ray Source
Perspective Projection (Ours)
Detector
X-ray features
Patient = Reconstructed CT using X-ray features
X-ray Source

Figure 2: (left) Orthogonal projection-based feature expansion. (right) Perspective projection-based feature extension.

2.2. Resampling Local Feature via the 2D-to-3D Module

The main role of the 2D-to-3D module M is relocating the coordinate-wise feature vector $\mathbf{f}_{l,i}^v$ extracted from X-ray local feature map $\mathbf{F}_l^v$ to a 3D CT space (Figure 1 (middle)). i is an index for 3D coordinates $\mathbf{x}_i$ on target CT slice, and detailed description is provided below.

$$\mathbf{f}_{l,i}^v = M(\mathbf{F}_l^v, \mathbf{x}_i). \tag{2}$$

This process is composed of two steps: 1) projecting the 3D coordinate $\mathbf{x}_i$ of a voxel into the corresponding 2D projection point $\mathbf{u}_i^v$ of each X-ray image plane, and 2) resampling the local feature vector $\mathbf{f}_{l,i}^v$ from local feature maps $\mathbf{F}_l^v$ of each X-ray image using the projection point $\mathbf{u}_i^v$.

Let $\{\mathbf{x}_i\}_{i=1}^{H_f\times W_f} \in \mathbb{R}^3$ be grid sampling points on the target CT slice $\mathbf{I}_n^a$, where n denotes the slice number of target CT, a is the imaging plane of CT ($a \in \{\text{axial}, \text{coronal}, \text{sagittal}\}$), and H_f, W_f the resolution of the resampled feature map.

For each point $\mathbf{x}_i$, we compute the corresponding projection point $\mathbf{u}_i^v \in \mathbb{R}^2$ in the image coordinates as follows:

$$\mathbf{u}_i^v = [\mathbf{R}^v(\theta)|\mathbf{t}^v]\,\mathbf{x}_i, \tag{3}$$

where $\mathbf{R}^v(\theta) \in \mathbb{R}^{3\times 3}$ is the rotation matrix and $\mathbf{t}^v \in \mathbb{R}^3$ is the translation vector, which are the extrinsic parameters of the X-ray source.

After that, we extract the feature vector $\mathbf{f}_{l,i}^v$ corresponding to the projection point $\mathbf{u}_i^v$ using bilinear interpolation:

$$\mathbf{f}_{l,i}^v = \mathbf{F}_l^v(\mathbf{u}_i^v). \tag{4}$$

The entire feature map is obtained by calculating $\mathbf{f}_{l,i}^v$ for all $\{\mathbf{x}_i\}_{i=1}^{H_f\times W_f}$ and is provided to the decoder as input.

	PSNR(↑)	SSIM(↑)	LPIPS(↓)	Params (M)	FLOPs (T)	Inference time (ms)
2DCNN [11]	25.398 ± 0.038	0.641 ± 0.001	0.483 ± 0.001	9.068	0.523	35.49 ± 1.73
X2CT-GAN-B [16]	26.013 ± 0.004	0.666 ± 0.003	0.349 ± 0.002	72.796	1.207	1560.87 ± 7.72
PerX2CT	$\mathbf{27.450 \pm 0.016}$	$\mathbf{0.732 \pm 0.000}$	$\mathbf{0.213 \pm 0.001}$	42.536	0.178	46.643 ± 1.43
PerX2CT$_{global}$	$\underline{27.335 \pm 0.004}$	$\underline{0.725 \pm 0.001}$	$\underline{0.216 \pm 0.003}$	70.131	0.190	52.82 ± 1.74

Table 1: Quantitative comparisons of CT reconstruction on the test set (mean ± std). Best and second best results are in **bold** and underlined.

Projection	PE	Global	Decoder	PSNR (↑)	SSIM (↑)	LPIPS (↓)
orthogonal	✓		2D	23.542 ± 0.118	0.576 ± 0.003	0.292 ± 0.003
perspective	✓		2D	$\mathbf{27.751 \pm 0.076}$	$\mathbf{0.749 \pm 0.003}$	$\mathbf{0.200 \pm 0.002}$
perspective	✓	✓	2D	$\underline{27.706 \pm 0.069}$	0.747 ± 0.003	$\underline{0.202 \pm 0.003}$
perspective			2D	27.685 ± 0.116	$\underline{0.748 \pm 0.005}$	$\mathbf{0.200 \pm 0.003}$
perspective			2D$_{mini}$	27.198 ± 0.175	0.727 ± 0.012	0.213 ± 0.008
perspective			3D$_{mini}$	26.294 ± 0.196	0.699 ± 0.007	0.302 ± 0.005

Table 2: Ablation study of PerX2CT on the validation set (mean ± std). Best and second best results are in **bold** and underlined.

		64 × 64			32 × 32		
PE	Global	PSNR (↑)	SSIM (↑)	LPIPS (↓)	PSNR (↑)	SSIM (↑)	LPIPS (↓)
		25.563 ± 0.557	0.793 ± 0.030	0.325 ± 0.020	23.969 ± 0.499	0.735 ± 0.037	0.341 ± 0.013
✓		$\underline{26.527 \pm 0.017}$	$\underline{0.841 \pm 0.001}$	**0.295 ± 0.001**	$\underline{25.029 \pm 0.052}$	$\underline{0.800 \pm 0.003}$	$\underline{0.314 \pm 0.001}$
✓	✓	**26.645 ± 0.096**	**0.846 ± 0.004**	$\underline{0.294 \pm 0.003}$	**25.701 ± 0.118**	**0.831 ± 0.004**	**0.307 ± 0.003**

Table 3: Ablation study of PerX2CT on the validation set for partial reconstruction (mean ± std). Best and second best results are in **bold** and underlined. We evaluate our model for 64 × 64 and 32 × 32 resolution depending on the size of the cropping part.

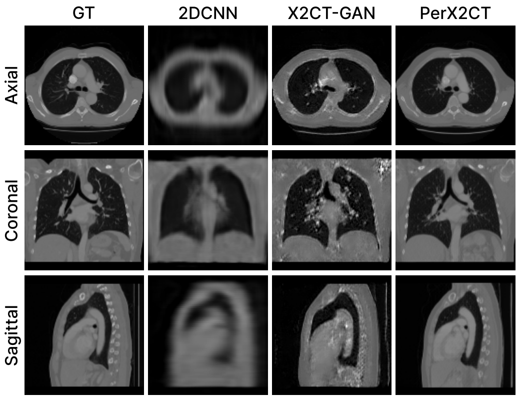

Figure 3: Qualitative results of CT reconstruction.

2.3. Decoding CT Slices

We independently reconstruct each CT slice using a 2D decoder instead of a 3D decoder for model efficiency. Our decoder takes five inputs: local feature maps $\{\mathbf{f}_{l,i}^{PA}\}_{i=1}^{H_f \times W_f}$, $\{\mathbf{f}_{l,i}^{Lat}\}_{i=1}^{H_f \times W_f}$, global feature maps $\mathbf{F}_g^{PA}, \mathbf{F}_g^{Lat}$, and the positional encoding of $\{\mathbf{x}_i\}_{i=1}^{H_f \times W_f}$. As global feature vectors $\mathbf{f}_g^v$ have a lower resolution than local feature maps, global feature vectors are spatially replicated to have the same spatial resolution as local feature maps by

$$\mathbf{f}_g^v \in \mathbb{R}^{C_g} \rightarrow \mathbf{F}_g^v \in \mathbb{R}^{H_f \times W_f \times C_g}. \tag{5}$$

We extract both local and global features on the two views and concatenate all of them channel-wise. The decoder $\mathbf{D}_\phi$ reconstructs the target slice $\mathbf{I}_n^a$ by

$$\hat{\mathbf{I}}_n^a = \mathbf{D}_\phi(\mathbf{F}_g^{PA}, \mathbf{F}_g^{Lat}, \{\mathbf{f}_{l,i}^{PA}, \mathbf{f}_{l,i}^{Lat}, \gamma(\mathbf{x}_i)\}_{i=1}^{H_f \times W_f}), \tag{6}$$

where $\gamma(\cdot)$ is the sinusoidal positional encoding [18], which maps each coordinate from $\mathbb{R}$ to $\mathbb{R}^{2L+3}$. To aggregate the global context, we use an attention layer for the lowest resolution feature map of the decoder.

2.4. Partial Reconstruction

Since the proposed method employs the coordinate-wise feature vector to reconstruct the CT slice, it can reconstruct the CT slice not only in full resolution (*i.e.*, full-frame) but also in arbitrarily selected patches. We added a cropped CT slice as training data to perform the partial reconstruction to train the model. Specifically, we randomly cropped the CT slice selected with the probability of p_{part} during the training phase. The cropping resolution is selected from the range from (H_{min}, W_{min}) to (H_{out}, W_{out}), and the bilinear interpolation is used. H_{min}, W_{min} are hyperparameters to determine the sampling range. If a cropped slice is given as a target, the 2D-to-3D module M samples the coordinate-wise local feature $\mathbf{f}_{l,i}^v$ within that part. Thus, PerX2CT can reconstruct the part of the CT slice in detail by densely sampling the corresponding features.

2.5. Overall Objective

Our overall objective function is given by

$$\mathcal{L}_{tot} = \lambda_{rec}\mathcal{L}_{rec} + \lambda_p\mathcal{L}_p, \tag{7}$$

where $\mathcal{L}_{rec}$ is the reconstruction loss which is the pixel-wise mean square error (MSE) for each slice, $\mathcal{L}_p$ is the perceptual loss [19]. λ_{rec}, λ_p control relative importance of the two losses.

3. EXPERIMENTS

3.1. Experimental Settings

Dataset. We require a dataset with X-ray and CT pairs to reconstruct CT from X-rays. Since collecting a real paired dataset is practically infeasible[1], existing studies have used digitally reconstructed radiograph (DRRs) technology to generate synthetic X-rays from a real CT volume [16]. We construct the dataset following Ying *et al.* [16], using LIDC-IDRI [20]. We randomly split into 916, 20, and 82 for training, validation, and test set. For CTs, we use slice images for the three planes (*i.e.*, axial, sagittal, and coronal) instead of full 3D data.

Evaluation Metrics. We use three different metrics to evaluate the quantitative quality of the predicted CT: peak signal-to-noise ratio (PSNR), structural similarity index (SSIM) [21], and learned perceptual image patch similarity (LPIPS) [19]. Since PSNR and SSIM do not properly reflect perspective quality [19, 22, 23], we use LPIPSs [19] which focuses on latent semantic perception and is highly correlated with human visual perception.

Implementation Details. For the image encoder, we fine-tune the ResNet-101[24] pre-trained on ImageNet as the backbone. The local feature map $\mathbf{F}_l^v$ and the global vector $\mathbf{f}_g^v$ are extracted from

[1] We need an X-ray pair (PA and Lateral) and a CT scan that were taken simultaneously, which hardly happens in reality.

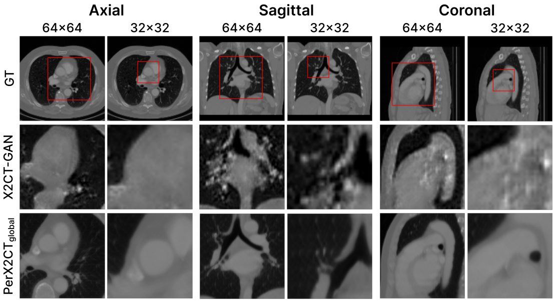

Figure 4: Qualitative results for partial reconstruction. The red bounding boxes denote the target patches for reconstruction, which have 64×64, and 32×32 resolutions.

the third and fourth layer blocks, respectively. The channel size of each feature C_l, C_g are 512, 256, and the spatial dimension of the local feature map $H_l \times W_l$ is 16×16. Our proposed model resamples the local feature map $\{\mathbf{f}_{l,i}^v\}_{i=1}^{H_f \times W_f}$, with set $H_f \times W_f$ as 32×32 resolution from $\mathbf{F}_l^v$. Without global feature, we use the encoder up to the third-layer block, and all other settings are the same. The decoder architecture follows [25], and set the number of layer blocks as 3. The output resolution of our decoder is 128 × 128. For positional encoding, we use L as 10. The weights in Eq. (7) are $\lambda_{rec} = \lambda_p = 1$. Each model train for 50 epochs (70 epochs for the partial reconstruction) with a batch size of 40. We use Adam optimizer [26] with an initial learning rate 4.5×10^{-6}. The momentum decay rate β_1 and the adaptive term decay rate β_2 are 0.5 and 0.9, respectively. We use Pytorch [27] on the RTX 3090 Ti. All experiments are repeated three times and the mean and standard deviation are reported. The code will be available at `https://github.com/dek924/PerX2CT`

3.2. Results

We compare the quantitative and qualitative results of our models with two baselines [11, 16]. Quantitative results are summarized in Table 1. PerX2CT only uses the perspective projection method and positional encoding, while PerX2CT$_{global}$ additionally uses the global latent vector. We show that the proposed methods have improvements over the baselines in the overall metrics, LPIPS, PSNR, and SSIM. We also compare the computational complexity of the models. We analyze it based on the number of trainable parameters, the number of the floating-point operation, and the inference speed. PerX2CT shows a 6.8× improvement in FLOPs and a 33.5× faster inference time compared to X2CT-GAN. 2DCNN is the best in terms of the number of parameters and inference speed, but FLOPs are about 3 times larger than PerX2CT and the quality of the reconstruction is significantly worse.

In Figure 3, we compare the visual quality of our proposed model and the baselines. Figure 3 shows that our reconstruction quality outperforms all baselines for all three views: axial, coronal and sagittal. 2DCNN reconstructs a blurrier image since it uses only one input X-ray. X2CT-GAN learns the boundary for large organs but fails to reconstruct accurately. PerX2CT successfully reconstructs the details of small anatomies, such as the atrium and aorta, clarifying the boundaries of the organs at the same time. Figure 4 illustrates the visual quality of the partial reconstruction of PerX2CT$_{global}$ and the baselines. We show our results for the selected part on the 64×64 and 32×32 resolution. The baseline results are cropping from the full-frame slice reconstructed by each model and then up-scaling to a full-frame resolution through bilinear interpolation. Unlike this, PerX2CT$_{global}$ directly reconstructs the target part without interpolation. We simply achieve that by adding randomly cropped data at training without any model architecture change.

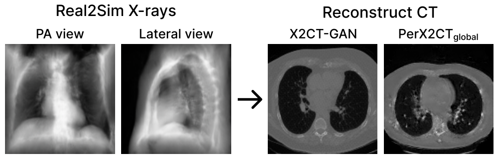

Figure 5: Qualitative results of the real X-ray projection.

3.3. Ablation Study

We analyze the effectiveness of our model components by comparing the performance of the four variants in the validation set. In rows 1-2 of Table 2, we evaluate the effectiveness of our feature expansion strategy, perspective projection, by showing a significant improvement for all metrics. We show the impact of positional encoding in both Table 2 and Table 3. PE also improves performance for all metrics. To clarify the benefit of our model architecture, we show the performance when all 2D convolutional layers in our model are converted to 3D. We conducted this experiment with a reduced number of layers, and without the attention layer due to memory issue for both 2D and 3D setting. As seen in rows 5-6 of the Table 2, 2D$_{mini}$ model outperforms the 3D$_{mini}$ model. The global features have an insignificant effect on reconstructing the full-frame CT slice (rows 2-3 of Table 2), but are effective for partial reconstruction (rows 2-3 in Table 3). The reason is that our decoder already aggregates the global context of the entire target CT slice through the self-attention layer. However, for partial reconstruction, the decoder receives only local features without additional global features.

3.4. Real-world Experiment

To obtain an X-ray pair (PA and Lateral) and a CT scan that were taken exactly at the same time is hardly happens in reality. Therefore, we trained Real2Sim CycleGAN [28] using 500 synthetic and real X-rays randomly selected from our training set and MIMIC-CXR [29][2], respectively. After that, we translated real X-rays by CycleGAN and used them as input of our model. Because there is no ground-truth CT corresponding to the X-ray, it is impossible to evaluate accurately. Instead, we provide qualitative results in Figure 5. As shown in Figure 5, PerX2CT provides a clearer boundary than in X2CT-GAN, even in a real-world setting.

4. CONCLUSION

In this paper, we propose PerX2CT, a perspective projection-based 3D CT reconstruction framework. PerX2CT utilizes 3D coordinate-dependent local feature extracted from the 2D biplanar X-ray feature, reflecting the perspective projection. We significantly improved the reconstruction performance of the full-frame slice while also reconstructing the desired part of the CT slice in a flexible manner. We validate that our proposed method outperforms existing models for all metrics. Our approach shows potential effectivenss in the medical field such as radiotherapy treatment planning with its high performance and being lightweight. For future work, we plan to expand our dataset to reflect more realistic situations, such as the presence of metal on X-rays.

[2] The MIMIC-CXR-JPG data were available on the project website at https://www.physionet.org/content/mimic-cxr-jpg/2.0.0/

5. REFERENCES

[1] David J. Brenner and Eric J. Hall, "Computed tomography–an increasing source of radiation exposure," *The New England journal of medicine*, pp. 2277–2284, 2007.

[2] Stephen Power, Fiachra Moloney, Maria Twomey, Karl James, Owen O'Connor, and Michael Maher, "Computed tomography and patient risk: Facts, perceptions and uncertainties," *World Journal of Radiology*, vol. 8, pp. 902, 12 2016.

[3] Charles Schmidt, "CT scans: Balancing health risks and medical benefits," *Environmental health perspectives*, vol. 120, pp. A118–21, 03 2012.

[4] H. Kudo et al., "Image reconstruction for sparse-view ct and interior ct-introduction to compressed sensing and differentiated backprojection," in *Quantitative imaging in medicine and surgery*, 2013.

[5] Dong Hye Ye, Gregery T. Buzzard, Max Ruby, and Charles A. Bouman, "Deep back projection for sparse-view CT reconstruction," in *GlobalSIP*, 2018.

[6] Weiwen Wu, Dianlin Hu, Chuang Niu, Hengyong Yu, Varut Vardhanabhuti, and Ge Wang, "Drone: Dual-domain residual-based optimization network for sparse-view ct reconstruction," *IEEE Transactions on Medical Imaging*, vol. 40, no. 11, pp. 3002–3014, 2021.

[7] Zhicheng Zhang, Xiaokun Liang, Xu Dong, Yaoqin Xie, and Guohua Cao, "A sparse-view ct reconstruction method based on combination of densenet and deconvolution," *IEEE Transactions on Medical Imaging*, vol. 37, pp. 1407–1417, 2018.

[8] Luke Pfister and Yoram Bresler, "Tomographic reconstruction with adaptive sparsifying transforms," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 6914–6918.

[9] Abril Corona-Figueroa, Jonathan Frawley, Sam Bond-Taylor, Sarath Bethapudi, Hubert P. H. Shum, and Chris G. Willcocks, "MedNeRF: Medical neural radiance fields for reconstructing 3d-aware ct-projections from a single x-ray," in *Proc. of the International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2022.

[10] Rongjun Ge, Yuting He, Cong Xia, Chenchu Xu, Weiya Sun, Guanyu Yang, Junru Li, Zhihua Wang, Hailing Yu, Daoqiang Zhang, et al., "X-ctrsnet: 3d cervical vertebra ct reconstruction and segmentation directly from 2d x-ray images," *Knowledge-Based Systems*, vol. 236, pp. 107680, 2022.

[11] Phlipp Henzler, Volker Rasch, Timo Ropinski, and Tobias Ritschel, "Single-image tomography: 3d volumes from 2d x-rays," *Computer Graphics Forum*, vol. 37, 10 2017.

[12] Ling Jiang, Mengxi Zhang, Ran Wei, Bo Liu, Xiangzhi Bai, and Fugen Zhou, "Reconstruction of 3d ct from a single x-ray projection view using cvae-gan," in *2021 IEEE International Conference on Medical Imaging Physics and Engineering (ICMIPE)*, 2021, pp. 1–6.

[13] Yoni Kasten, Daniel Doktofsky, and Ilya Kovler, "End-to-end convolutional neural network for 3d reconstruction of knee bones from bi-planar x-ray images," in *International Workshop on Machine Learning for Medical Image Reconstruction*, 2020.

[14] Md Aminur Rab Ratul, Kun Yuan, and WonSook Lee, "Ccx-raynet: A class conditioned convolutional neural network for biplanar x-rays to ct volume," in *IEEE International Symposium on Biomedical Imaging (ISBI)*, 2021.

[15] Liyue Shen, Wei Zhao, and Lei Xing, "Patient-specific reconstruction of volumetric computed tomography images from a single projection view via deep learning," *Nature Biomedical Engineering*, vol. 3, 11 2019.

[16] Heng Guo Xingde Ying, Jian Wu Kai Ma, Zhengxin Weng, and Yefeng Zheng, "X2ct-gan: reconstructing ct from biplanar x-rays with generative adversarial networks," in *CVPR*, 2019.

[17] Tomaso A. Poggio Mario Bertero and Vincent Torre, "Ill-posed problems in early vision," *Proc. of the IEEE*, vol. 76, no. 8, pp. 869–889, 1988.

[18] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng, "NeRF: Representing scenes as neural radiance fields for view synthesis," in *ECCV*, 2020.

[19] Phillip Richard, Eli Alexei, and Oliver, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018.

[20] McLennan Armato, McNitt-Gray Bidaut L., and *et al*. Meyer. C. R., "The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans," in *Medical physics*, 2011.

[21] Bovik A. C. Wang Z. and Simoncelli E. P. Sheikh H. R., "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process*, p. 600–612, 2004.

[22] Daeho Lee and Sungsoo Lim, "Improved structural similarity metric for the visible quality measurement of images," *Journal of Electronic Imaging*, vol. 25, 12 2016.

[23] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[24] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," *CVPR*, pp. 770–778, 2016.

[25] Björn Ommer Patrick Esser, Robin Rombach, "Taming transformers for high-resolution image synthesis," in *CVPR*, 2021.

[26] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

[27] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *NeurIPS*, 2019.

[28] Jun-Yan Zhu et al., "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *ICCV*, 2017.

[29] Johnson et al., "Mimic-cxr: A large publicly available database of labeled chest radiographs," in *arXiv*, 2019.