C^2RV: Cross-Regional and Cross-View Learning for Sparse-View CBCT Reconstruction

Yiqun Lin[1] Jiewen Yang[1] Hualiang Wang[1] Xinpeng Ding[1] Wei Zhao[2*] Xiaomeng Li[1*]
[1]The Hong Kong University of Science and Technology [2]Beihang University
{ylindw, jyangcu, hwangfd, xdingaf}@connect.ust.hk,
zhaow20@buaa.edu.cn, eexmli@ust.hk

Abstract

Cone beam computed tomography (CBCT) is an important imaging technology widely used in medical scenarios, such as diagnosis and preoperative planning. Using fewer projection views to reconstruct CT, also known as sparse-view reconstruction, can reduce ionizing radiation and further benefit interventional radiology. Compared with sparse-view reconstruction for traditional parallel/fan-beam CT, CBCT reconstruction is more challenging due to the increased dimensionality caused by the measurement process based on cone-shaped X-ray beams. As a 2D-to-3D reconstruction problem, although implicit neural representations have been introduced to enable efficient training, only local features are considered and different views are processed equally in previous works, resulting in spatial inconsistency and poor performance on complicated anatomies. To this end, we propose C^2RV by leveraging explicit multi-scale volumetric representations to enable cross-regional learning in the 3D space. Additionally, the scale-view cross-attention module is introduced to adaptively aggregate multi-scale and multi-view features. Extensive experiments demonstrate that our C^2RV achieves consistent and significant improvement over previous state-of-the-art methods on datasets with diverse anatomy. Code is available at https://github.com/xmed-lab/C2RV-CBCT.

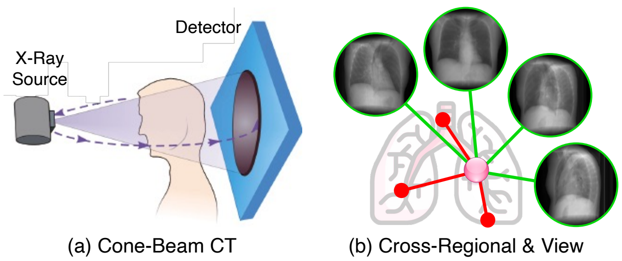

Figure 1. (a) Cone-shaped X-ray beams are emitted from the scanning source and a 2D array of detectors measures the transmitted radiation. (b) Cross-regional (red) and cross-view (green) feature learning to enhance point-wise representation.

1. Introduction

Computed tomography (CT) has become an indispensable technique used for medical diagnostics, providing accurate and non-invasive visualization of internal anatomical structures. Compared with conventional CT (fan/parallel-beam), cone-beam CT (CBCT) offers advantages, including faster acquisition and improved spatial resolution [28]. Typically, hundreds of projections are required to produce a high-quality CT scan involving high radiation doses from X-rays. However, high radiation dose exposure to patients can be a concern in clinical practice, limiting its use in scenarios like interventional radiology. Hence, reducing the number of projections can be one of the ways to reduce the radiation doses, which is also known as sparse-view reconstruction.

Over the past decades, there have been many research works studying the sparse-view problem for conventional CT by formulating the reconstruction as a mapping from 1D projections to a 2D CT slice, where generation-based techniques [6, 7, 10, 13, 20, 20, 35, 37, 45] are proposed to operate on the image or projection domains. However, the measurements of cone-beam CT are 2D projections (Figure 1a), resulting in increased dimensionality compared with conventional CT. This means that extending previous conventional CT reconstruction methods to CBCT will encounter issues [18] such as high computational cost.

Recently, implicit neural representations (INRs) have been widely used in 3D reconstruction, including novel view synthesis and object reconstruction. To handle sparse-view or even single-view scenarios, geometric priors (*e.g.*, surface points [40] and normals [41]) or parametric shape models [11, 38, 39, 46] (*e.g.*, SMPL [19] and SMPL-X [24]) are incorporated to improve the robustness and generalization ability. However, unlike visible light, X-rays have a

*Corresponding Authors

higher frequency and pass through the surfaces of many materials, hence, no depth or surface information can be measured in the projection. Additionally, it is difficult to build a CT-specific parametric model as the internal anatomies of the human body are more complicated than surface models.

Although INRs have been introduced to CBCT reconstruction in recent years, tens of views (*i.e.*, 20-50) are still required for self-supervised NeRF-based methods [3, 31, 44] due to the lack of prior knowledge. On the other hand, current data-driven methods like DIF-Net [18] may suffer from poor performance when the anatomy has complicated structures for two possible reasons: 1.) local features queried from projections can be difficult to identify different organs that have low contrast in the projection; 2.) projections of different views are processed equally, while some views indeed present more information of specific organs than other views. For example, the right-left view shows the patella clearly, while it overlaps the femur in the anterior-posterior view; see Figure 2.

To address the limitations of previous works, we propose a novel sparse-view CBCT reconstruction framework C^2RV by leveraging cross-regional and cross-view feature learning to enhance point-wise representation (Figure 1b). To be more specific, we first introduce multi-scale 3D volumetric representations (MS-3DV), where features are obtained by back-projecting multi-view features at different scales to the 3D space. Explicit MS-3DV enables cross-regional learning in 3D space, providing richer information that helps better identify different organs. Hence, the feature of a point can be queried in a hybrid way, *i.e.*, multi-scale voxel-aligned features from MS-3DV and multi-view pixel-aligned features from projections. Instead of considering queried features equally, scale-view cross-attention (SVC-Att) is then proposed to adaptively learn aggregation weights by self-attention and cross-attention. Finally, multi-scale and multi-view features are aggregated to estimate the attenuation coefficient. We evaluate C^2RV quantitatively and qualitatively on two CT datasets (*i.e.*, chest and knee). Extensive experiments demonstrate that our proposed C^2RV consistently outperforms previous state-of-the-art methods by a considerable margin under different experimental settings.

The main contributions of this work are summarized as:

- Multi-scale 3D volumetric representations (MS-3DV) to enable cross-regional learning in the 3D space;
- Scale-view cross-attention (SVC-Att) to adaptively aggregate multi-scale and multi-view features;
- C^2RV, a novel sparse-view CBCT reconstruction framework, achieving state-of-the-art performance on datasets with diverse anatomy.
- Ablative studies to analyze the effectiveness and robustness of the proposed C^2RV.

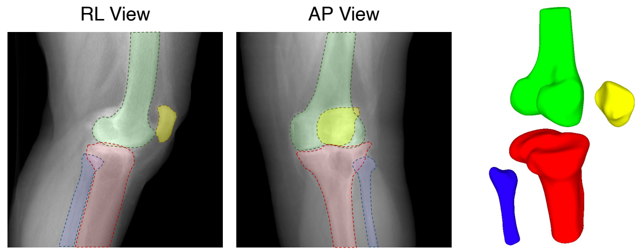

Figure 2. Right-left (RL) and anterior-posterior (AP) views of the knee. Green: femur. Red: tibia. Yellow: patella. Blue: fibula. The patella and femur overlap in the AP view but not in the RL view.

2. Related Work

In computer vision, especially 3D vision, the reconstruction problem has gained significant attention in recent years. In this section, we mainly review related work of sparse-view reconstruction on traditional parallel/fan-beam CT, cone-beam CT, and general 3D.

2.1. Sparse-View CT Reconstruction

Traditional parallel/fan-beam CT reconstruction can be regarded as reconstructing a 2D CT slice from 1D projections. Existing learning-based methods mainly include image-domain, projection-domain, and dual-domain methods. Specifically, image-domain methods [6, 10, 13, 20, 35, 45] apply filtered back projection (FBP) to reconstruct a coarse CT slice with streak artifacts and utilize CNNs, such as U-Net [25] and DenseNet [9], to denoise and refine details. When extending these methods to CBCT reconstruction, the network should be modified to 3D CNNs, resulting in a substantial increase in computational cost. Another way is to adopt these methods for slice-wise (2D) denoising [15], while the 3D spatial consistency cannot be guaranteed.

Projection-domain methods directly operate on sparse-view 1D projections by mapping the projections to the CT slice [7] or recovering the full-view projections [37]. Additionally, Song *et al.* [32] utilize score-based generative models and propose a sampling method to reconstruct an image consistent with both the measurement process and the observed measurements (*i.e.*, projections). Chung *et al.* [2] further incorporate 2D diffusion models into iterative reconstruction. Dual-domain methods operate on both projection and image domains by combining the denoising processes of two domains [17, 20] or modeling dual-domain consistency [34]. However, projection-based operations cannot be extended to CBCT reconstruction as the measurement processes (cone-beam *vs*. parallel/fan-beam) are different.

2.2. Sparse-View CBCT Reconstruction

Different from traditional parallel/fan-beam CT, the measurement of cone-beam CT is a 2D projection, which means the reconstruction should be formulated as reconstructing a

3D CT volume from multiple 2D projections. Conventional filtered back-projection (FDK [4]) and ART-based iterative methods [1, 5, 22] often suffer from heavy streaking artifacts and poor image quality when the number of projections is dramatically decreased. Recently, learning-based approaches are proposed for single/orthogonal-view CBCT reconstruction [12, 14, 30, 42], while these methods are specially designed for single/orthogonal-view reconstruction [12, 14, 42] or patient-specific data [30], making them difficult to extend to general sparse-view reconstruction.

On the other hand, implicit neural representations [21, 26] have been introduced to represent CBCT as an attenuation [3, 44] or intensity [18] field. Self-supervised methods, including NAF [44] and NeRP [31], simulate the measurement process and minimize the error between real and synthesized projections. However, these methods require a long time for per-sample optimization and are only suitable for the reconstruction from tens of views (*i.e.*, 20-50) due to the lack of prior knowledge. DIF-Net [18], as a data-driven method, formulates the problem as learning a mapping from sparse projections to the intensity field. Nevertheless, DIF-Net regards different projections equally, and only local semantic features are queried for each sampled point, leading to limited reconstruction quality when processing anatomies with complicated structures (*e.g.*, chest).

2.3. Sparse-View 3D Reconstruction

In 3D computer vision, implicit representations have been widely used in novel-view synthesis [21, 40, 41, 43] and object reconstruction [11, 23, 27, 38, 39, 46]. For novel view synthesis, to extend NeRF [21] to sparse-view scenarios, geometric priors like surface points [40] and normals [41] are incorporated to improve the generalization ability and efficiency. For object reconstruction, particularly digital human reconstruction, previous works [11, 38, 39, 46] leverage explicit parametric SMPL(-X) [19, 24] models to constrain surface reconstruction and improve the robustness. However, there is no available depth or surface information in the attenuation fields of CBCT since X-rays penetrate right through many common materials, such as flesh. SMPL(-X) are 3D parametric shape models specially designed for the surface of the human body, while the internal anatomy structures are too complicated to design a CT-specific parametric model. Therefore, parametric shape models cannot be used in sparse-view CBCT reconstruction. Furthermore, cross-view relationships are rarely considered in surface-based reconstruction since one or two views are more practical and often sufficient to learn the sparse field with the above-mentioned priors.

3. Methodology

In this section, we first revisit the problem formulation of sparse-view CBCT reconstruction and the baseline DIF-Net proposed in [18]. Then, we formally introduce C^{2}RV, consisting of multi-scale 3D volumetric representations (MS-3DV) and the scale-view cross-attention (SVC-Att) for cross-regional and cross-view learning.

3.1. Revisit DIF-Net [18]

We follow previous works [18, 44] to formulate the CT image as a continuous implicit function $g\colon \mathbb{R}^3 \rightarrow \mathbb{R}$, which defines the attenuation coefficient (same as "intensity" in [18]) $v \in \mathbb{R}$ of a point $p \in \mathbb{R}^3$ in the 3D space, *i.e.*, $v = g(p)$. Hence, given N-view projections $\mathcal{I} = \{I_1, \ldots, I_N\} \subset \mathbb{R}^{W \times H}$ (W and H are width and height) with known scanning parameters (*e.g.*, viewing angles, distance of source to origin) during the measurement process, the reconstruction problem is formulated as a conditioned implicit function $g(\cdot)$ such that $v = g(\mathcal{I}, p)$.

In practice, a 2D encoder-decoder (shared across different views) is used to extract multi-view feature maps $\mathcal{F} = \{\mathcal{F}_1, \ldots, \mathcal{F}_N\} \subset \mathbb{R}^{C \times (W \times H)}$ from N-view projections $\mathcal{I}$, where C is the output channel size of the decoder. For i^{th} view, denote the projection function as $\pi_i\colon \mathbb{R}^3 \rightarrow \mathbb{R}^2$, which maps a 3D point p to the 2D plane where detectors are located such that $p'_i = \pi_i(p)$. Then, we define the view-specific pixel-aligned features of p in i^{th} view as

$$
\begin{aligned}
\mathcal{F}_i(p) &= \text{Interp}\big(\mathcal{F}_i, \pi_i(p)\big) \\
&= \text{Interp}(\mathcal{F}_i, p'),
\end{aligned} \tag{1}
$$

where Interp: $(\mathbb{R}^{C \times (D_1 \times \cdots \times D_k)}, \mathbb{R}^k) \rightarrow \mathbb{R}^C$ is k-linear interpolation. Particularly, $k = 2$ and Interp$(\cdot)$ is bilinear interpolation in the above equation.

Denoting multi-view pixel-aligned features of p as $\mathcal{F}(p) = \big\{\mathcal{F}_1(p), \ldots, \mathcal{F}_N(p)\big\} \subset \mathbb{R}^C$, the attenuation coefficient of p is

$$
v = g(\mathcal{I}, p) = \sigma\big(\mathcal{F}(p)\big), \tag{2}
$$

where $\sigma(\cdot)$ is the aggregation function implemented with MLPs (or Max-Pooling + MLPs) in DIF-Net [18]. Although the above formulation and implementation enable efficient training for high-resolution sparse-view reconstruction, only local pixel-aligned features queried from projections are considered and different views are processed equally, leading to poor performance on complicated anatomies; see analysis in Sec. 1 & 2.2 and results in Table 1. To this end, we propose C^{2}RV and will introduce it in detail in the following section.

3.2. C^{2}RV Framework

C^{2}RV (**C**ross-**R**egional and **C**ross-**V**iew Learning) framework is developed based on DIF-Net [18] to address the above-mentioned limitations. The framework overview is shown in Figure 3. Specifically, multi-scale 3D volumetric representations (MS-3DV) are obtained by back-projecting

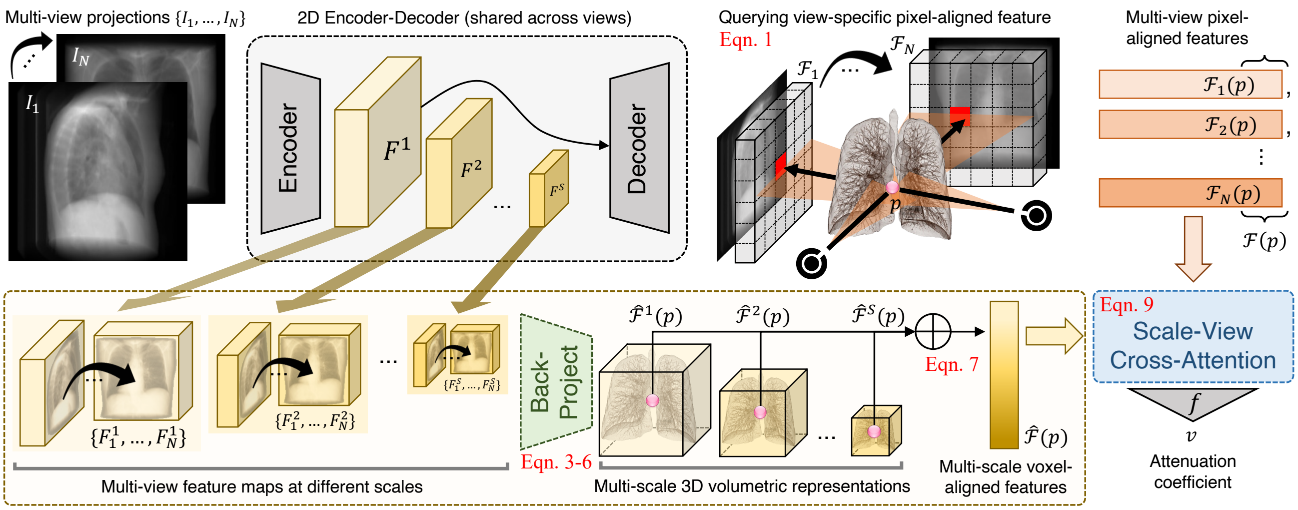

Figure 3. The overview of the proposed sparse-view reconstruction framework C^{2}RV. Given multi-view projections, a 2D encoder-decoder is applied to extract view-wise feature map $\mathcal{F}_i$ for querying the pixel-aligned feature $\mathcal{F}_i(p)$. Additionally, the output feature map F^1 of the encoder is downsampled to obtain multi-scale feature maps. At each scale s, multi-view features are back-projected to the 3D space and gathered to form the 3D volumetric representation $\hat{\mathcal{F}}^s$ for querying the voxel-aligned feature $\hat{\mathcal{F}}^s(p)$. Finally, multi-scale voxel-aligned features and multi-view pixel-aligned features are aggregated via scale-view cross-attention modules to estimate the attenuation coefficient.

multi-view feature maps at different scales to the 3D space. Hence, multi-scale voxel-aligned features and multi-view pixel-aligned features are adaptively aggregated by scale-view cross-attention (SVC-Att) modules to estimate the attenuation coefficient.

Low-Resolution 3D Volumetric Representation. A 3D volumetric space $\mathcal{S} \in \mathbb{R}^{3\times(r\times r\times r)}$ is defined by voxelizing the 3D space with a low resolution $r \leq 16$. Let $F_i \in \mathbb{R}^{c\times w\times h}$ be the intermediate feature map of the encoder-decoder given the projection of i^{th} view. The volumetric feature space $\hat{F} \in \mathbb{R}^{c\times(r\times r\times r)}$ defined over $\mathcal{S}$ is produced by back-projecting multi-view feature maps into $\mathcal{S}$, *i.e.*,

$$\hat{F} = \text{Back-Project}\big(\{F_1, \ldots, F_N\}, \mathcal{S}\big), \tag{3}$$

where the feature of a voxel q in $\mathcal{S}$ is

$$\begin{aligned} \hat{F}(q) &= \varphi\Big(\big\{F_1(q), \ldots, F_N(q)\big\}\Big), \\ \text{where } F_i(q) &= \text{Interp}\big(F_i, \pi_i(q)\big), \end{aligned} \tag{4}$$

and $\varphi(\cdot)$ is the aggregation function, implemented with Max-Pooling in practice. Therefore, 3D convolutional layers (denoted as ϕ) can be followed for efficient cross-regional feature learning, *i.e.*,

$$\hat{\mathcal{F}} = \phi(\hat{F}). \tag{5}$$

MS-3DV: Multi-Scale 3D Volumetric Representations. To further improve the robustness of reconstructing different anatomical structures, we propose to leverage multi-scale 3D volumetric representations. To be specific, given the projection of i^{i} view, denote the output feature map of the encoder as F_i^1, then a sequence of downsampling operators ρ are applied to produce multi-scale feature maps $\{F_i^1, \ldots, F_i^S\}$, where $F_i^s = \rho_{s-1}(F_i^{s-1})$ for $s \in \{2, \ldots, S\}$, and S is the total number of scales. Then, we define multi-scale 3D voxelized space $\{\mathcal{S}^1, \cdots, \mathcal{S}^S\}$ with different resolutions $\{r^1, \ldots, r^S\}$, and back-project (Eqn. 3 and 5) multi-view feature maps of each scale to obtain multi-scale 3D volumetric representations (MS-3DV) $\{\hat{\mathcal{F}}^1, \ldots, \hat{\mathcal{F}}^S\}$, where

$$\hat{\mathcal{F}}^s = \phi^s\Big(\text{Back-Project}\big(\{F_1^s, \ldots, F_N^s\}, \mathcal{S}^s\big)\Big), \tag{6}$$

for $s \in \{1, \ldots, S\}$. Hence, in addition to multi-view pixel-aligned features directly queried from view-specific feature maps, we incorporate multi-scale voxel-aligned features for the point p into the estimation of the attenuation coefficient, as given by

$$\hat{\mathcal{F}}(p) = \text{MLPs}\Big(\text{Concat}\big[\hat{\mathcal{F}}^1(p), \ldots, \hat{\mathcal{F}}^S(p)\big]\Big), \tag{7}$$

where $\hat{\mathcal{F}}^s(p) = \text{Interp}(\hat{\mathcal{F}}^s, p)$, Concat$[\cdot]$ indicates concatenation, and multi-layer perceptrons (MLPs) map the channel size of concatenated voxel-aligned features to be consistent with pixel-aligned features (Eqn. 1), *i.e.*, C.

SVC-Att: Scale-View Cross-Attention. We first recall the definition of cross-attention (C-Att) [33] given the reference features $F_r \in \mathbb{R}^{L_r\times C_r}$ and source features $F_s \in \mathbb{R}^{L_s\times C_s}$,

$$\begin{aligned} \text{C-Att}(F_r, F_s) &= \text{softmax}\Big(\frac{QK^T}{\sqrt{C_d}}\Big)V, \\ \text{where } Q = F_rM_q,\ & K = F_sM_k,\ V = F_sM_v, \end{aligned} \tag{8}$$

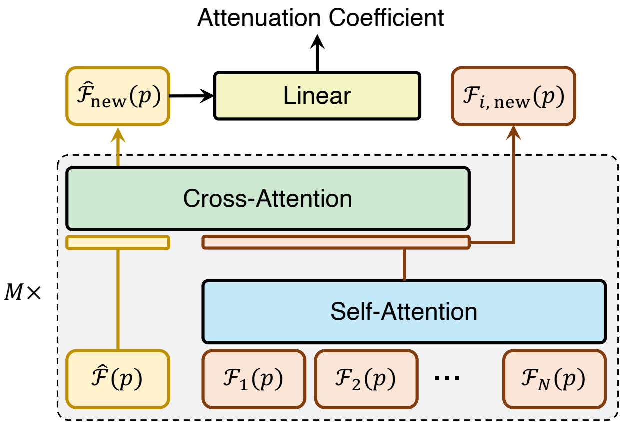

Figure 4. The overview of scale-view cross attention (SVC-Att) module. In each SVC-Att module, a self-attention is first applied to multi-view features, and then a cross-attention is followed to conduct attention between multi-scale features and multi-view features. M SVC-Att modules are stacked and finally followed by a linear layer to estimate the attenuation coefficient.

and $M_q \in \mathbb{R}^{C_r \times C_d}$, $M_k, M_v \in \mathbb{R}^{C_s \times C_d}$. Self-attention (S-Att) can be regarded as a special case of cross-attention where we let $F_s = F_r$.

For a point p, let $\mathcal{F}(p) = \{\mathcal{F}_1(p), \ldots, \mathcal{F}_N(p)\} \subset \mathbb{R}^C$ denote multi-view pixel-aligned features queried from projections (Eqn. 1), and $\hat{\mathcal{F}}(p) \in \mathbb{R}^C$ indicate the multi-scale voxel-aligned features queried from MS-3DV (Eqn. 7). The scale-view cross-attention (SVC-Att) module is proposed to adaptively aggregate the above features. As shown in Figure 4, a self-attention module is first applied to conduct cross-view attention on multi-view features $\mathcal{F}(p)$. Then, a cross-attention attention module takes multi-scale features as the reference and the output of the self-attention module as the source to conduct attention between multi-scale and multi-view features. To formulate,

$$\begin{aligned} \hat{\mathcal{F}}_{\text{new}}(p) &= \text{SVC-Att}\big(\hat{\mathcal{F}}(p), \mathcal{F}(p)\big) \\ &= \text{C-Att}\big(\hat{\mathcal{F}}(p), \mathcal{F}_{\text{new}}(p)\big), \\ \text{where } \mathcal{F}_{\text{new}}(p) &= \text{S-Att}\big(\mathcal{F}(p)\big). \end{aligned} \tag{9}$$

In practice, M SVC-Att modules are stacked and a linear layer is followed to estimate the attenuation coefficient.

3.3. Network Training

We follow [18] to train the reconstruction network on a CT dataset, where the projections are simulated from the CT by digitally reconstructed radiographs (DRRs). Specifically, we denote the volumetric CT as $I_{\text{ct}} \in \mathbb{R}^{1 \times (W_{\text{ct}} \times H_{\text{ct}} \times D_{\text{ct}})}$ and the projections as $\mathcal{I}$. Then, the ground-truth attenuation field defined over the continuous 3D space $\mathcal{P}$ is

$$\mathcal{V} = \big\{v(p) = \text{Interp}(I_{\text{ct}}, p) \;\big|\; \forall p \in \mathcal{P}\big\}, \tag{10}$$

where $\text{Interp}(\cdot)$ is the interpolation operator (Eqn. 1). The estimated attenuation field by C^{2}RV is given as

$$\hat{\mathcal{V}} = \big\{\hat{v}(p) = g(\mathcal{I}, p) \;\big|\; \forall p \in \mathcal{P}\big\}. \tag{11}$$

Hence, the mean square error (MSE) as the objective function is used to compute point-wise estimation error,

$$\mathcal{L}_{\text{MSE}}\big(\mathcal{V}, \hat{\mathcal{V}}\big) = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \Big(v(p) - \hat{v}(p)\Big)^2. \tag{12}$$

During each training iteration, we randomly sample 10,000 points from $\mathcal{P}$ for loss calculation (Eqn. 12) to reduce the memory requirements for efficient network optimization. During the inference, the 3D space is voxelized with a specified resolution (*e.g.*, 256^3), where the attenuation coefficient of a voxel is defined as the estimated attenuation coefficient of its centroid point by C^{2}RV. This means that the resolution can be chosen based on the desired trade-off between image quality and reconstruction speed.

Implementation. In practice, we empirically choose $S = 3$, $r^1 = 16$, and $r^s = \frac{1}{2} r^{s-1}$ for $s \geq 2$. We follow [18] to use U-Net [25] with $C = 128$ output feature channels as the 2D encoder-decoder, where the size of encoder output F^1 is $\frac{W}{16} \times \frac{H}{16}$. $\phi(\cdot)$ in Eqn. 5 is implemented with 3-layer 3D residual convolution that maps the channel size of $\hat{F}$ to C. For the aggregation method, $M = 3$ SVC-Att modules are stacked, and attention modules are implemented as multi-head attention with 8 heads. During training, the learnable parameters of C^{2}RV are optimized using stochastic gradient descent (SGD) with a momentum of 0.98 and an initial learning rate of 0.01. We train C^{2}RV with 400 epochs and a batch size of 4. The learning rate is decreased by a factor of $(10^{-3})^{1/400}$ per epoch.

4. Experiments

To validate the effectiveness of our proposed C^{2}RV, we conduct experiments on two CT datasets with different anatomies, including chest and knee. In addition to quantitative and qualitative evaluation, automatic segmentation is applied to sparse-view reconstruction results, showing the practical potential of reconstructed CT by C^{2}RV in downstream applications.

4.1. Experimental Setting

Dataset. Experiments are conducted on two CT datasets, including a public chest CT dataset (LUNA16 [29]) and a private knee CBCT dataset collected by Lin *et al.* [18] (additional experiments on a dental CBCT dataset are provided in the supplementary). Specifically, LUNA16 [29] is composed of 888 chest CT scans with resolution ranging from $145 \times 145 \times 108$ to $375 \times 375 \times 509$ mm^3, split into 738 for training, 50 for validation, and 100 for testing; the knee

Table 1. Comparison of different methods on two CT datasets (*i.e.*, chest and knee) with various numbers of projection views. The resolution of the reconstructed CT is 256^3. The reconstruction results are evaluated with PSNR (dB) and SSIM ($\times 10^{-2}$), where higher PSNR/SSIM indicate better performance. The best values are **bolded** and the second-best values are underlined.

Method	Type	LUNA16 [29] (Chest CT)			Lin *et al.* [18] (Knee CBCT)		
		6-View	8-View	10-View	6-View	8-View	10-View
FDK [4]	Self-Supervised	15.34|35.78	16.58|37.89	17.40|39.85	17.71|37.49	19.23|40.51	20.50|43.64
SART [1]		19.70|64.36	20.06|67.80	20.23|70.23	24.73|80.71	25.81|84.08	26.72|86.15
NAF [44]		18.76|54.16	20.51|60.84	22.17|62.22	20.11|58.43	22.42|67.19	24.26|75.02
NeRP [31]		23.55|74.46	25.83|80.67	26.12|81.30	24.24|70.05	25.55|74.68	26.33|79.81
FBPConvNet [13]	Data-Driven: Denoising	24.38|77.57	24.87|78.86	25.90|80.03	25.10|83.35	25.93|83.47	26.74|84.46
FreeSeed [20]		25.59|77.36	26.86|78.92	27.23|79.25	26.74|84.19	27.88|85.62	28.77|87.04
BBDM [16]		24.78|77.03	25.81|78.06	26.35|79.38	26.58|84.33	28.01|85.46	28.90|87.25
PixelNeRF [43]	Data-Driven: INR-based	24.66|78.68	25.04|80.57	25.39|82.13	26.10|87.69	26.84|88.75	27.36|89.58
DIF-Net [18]		25.55|84.40	26.09|85.07	26.67|86.09	27.12|89.12	28.31|90.24	29.33|92.06
C^{2}RV (*ours*)		**29.23|92.78**	**29.95|93.49**	**30.70|94.03**	**29.73|93.64**	**30.68|94.42**	**31.55|95.01**

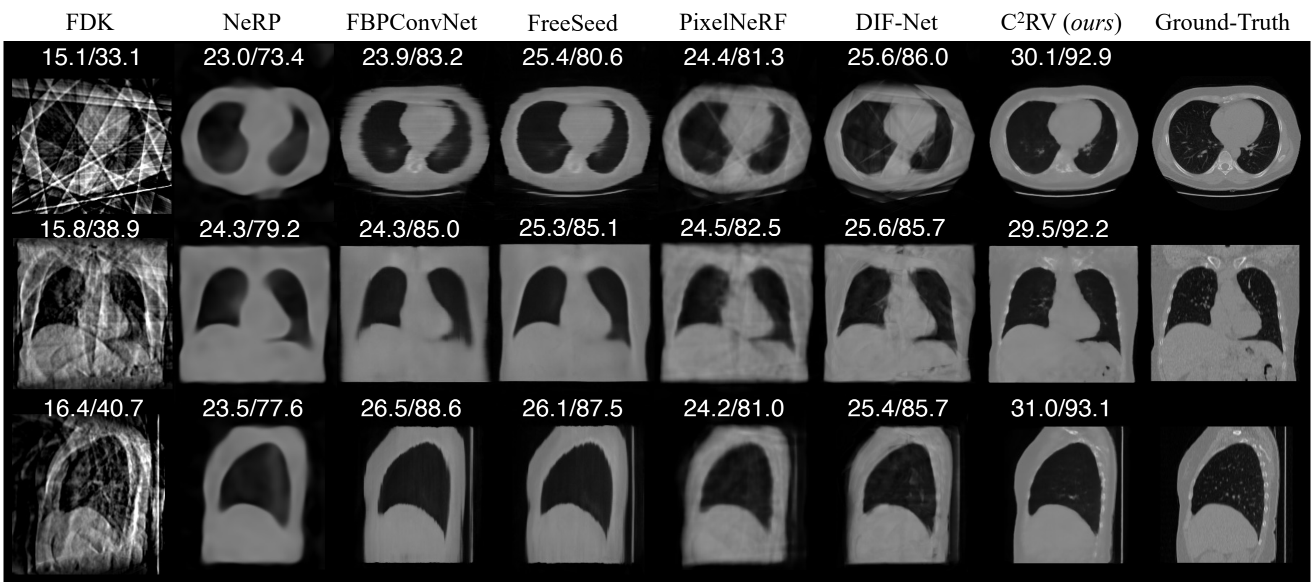

Figure 5. Visualization of 6-view reconstructed chest CT (from top to bottom: axial, coronal, and sagittal slice). PSNR/SSIM (dB/$\times 10^{-2}$) values are presented above each visualized example.

dataset [18] contains 614 knee CBCT scans with resolutions ranging from $236\times236\times167$ to $500\times500\times416$ mm^3, split into 464 for training, 50 for validation, and 100 for testing. We follow the data preprocessing of [18] to resample and crop (or pad) each CT to have isotropic spacing (*i.e.*, 1.6 mm for chest and 0.8 mm for knee) and size of 256^3. Multi-view 2D projections are simulated by DRRs with a resolution of 256^2, and the viewing angles are uniformly selected in the range of 180° (half rotation).

Evaluation Metrics. Following previous works [18, 31, 44], two quantitative metrics, including peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) [36], are used to evaluate the reconstruction performance, where higher values indicate superior image quality.

4.2. Results

Quantitative Evaluation. We compare our C^{2}RV with self-supervised methods, including FDK [4], SART [1], NAF [44], and NeRP [31], without requiring additional training data. We also compare data-driven approaches, including 2D denoising-based (*i.e.*, FBPConvNet [13], FreeSeed [20], and BBDM [16]) and implicit neural representation (INR)-based (*i.e.*, PixelNeRF [43] and DIF-Net [18]) methods. We conduct experiments with different numbers of projection views (*i.e.*, 6-10) and the reconstruction resolution is 256^3. The results are shown in Table 1. Although DIF-Net [18] can achieve satisfactory performance on knee CT, the performance drops dramatically when adapting to more complicated anatomical structures

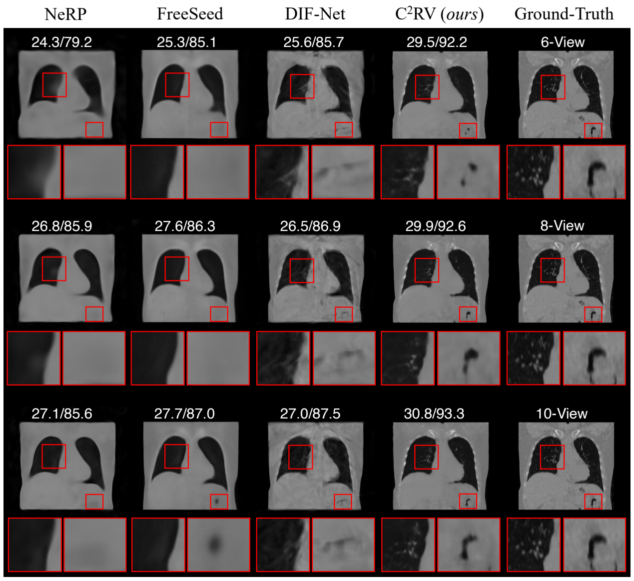

Figure 6. Visualization of examples reconstructed from different numbers of projection views, *i.e.*, 6, 8, and 10. The highlighted regions (red) are zoomed in, showing richer details in our reconstructed results than in other methods.

Table 2. Lung segmentation of 6-view reconstructed chest CT. Dice coefficient (%, higher is better) and average surface distance (ASD, mm, lower is better) are evaluated. The best values are **bolded** and the second-best values are <u>underlined</u>.

Method	Recon.		Left Lung		Right Lung	
	PSNR	SSIM	Dice	ASD↓	Dice	ASD↓
FDK [4]	15.34	35.78	16.51	79.55	46.14	22.44
NeRP [31]	23.55	74.46	86.55	9.57	86.24	3.62
FBPConvNet [13]	24.38	77.36	92.78	3.14	91.37	2.68
FreeSeed [20]	<u>25.59</u>	77.36	<u>95.16</u>	<u>1.74</u>	94.75	<u>1.77</u>
PixelNeRF [43]	24.66	78.68	91.00	5.31	91.66	3.67
DIF-Net [18]	25.55	<u>84.40</u>	94.45	2.51	<u>94.78</u>	2.01
C^{2}RV (*ours*)	**29.23**	**92.78**	**96.72**	**1.25**	**96.93**	**1.12**

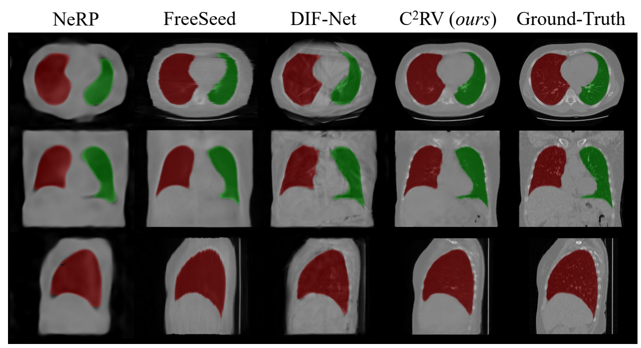

Figure 7. Visualization of lung segmentation on 6-view reconstructed chest CT. Red: left lung. Green: right lung.

(*e.g.*, chest), while our C^{2}RV consistently performs well on different datasets. Additionally, when reconstructing from 6, 8, and 10 views, our C^{2}RV outperforms previous state-of-the-art by a remarkable margin, *i.e.*, 3.6/8.4, 3.1/8.4, and 3.5/7.9 PSNR/SSIM (dB/$\times 10^{-2}$) on chest CT; and 2.6/4.5, 2.4/4.2, and 2.2/3.0 on knee CT. More importantly, Even with only 6 views, C^{2}RV can reconstruct CT of better quality than other methods with 4 more views (*i.e.*, 10 views).

Visual Comparison. Examples of 6-view reconstruction are visualized in Figure 5 for qualitative comparison. Due to the lack of sufficient projection views, reconstruction results of FDK [4] are full of streaking artifacts, and NeRP [31] can only reconstruct satisfactory contours of the body and lung. For FBPConvNet [13] and FreeSeed [20], jitters appear near the boundary of the body and lung since they are 2D methods that reconstruct CT slice by slice. For PixelNeRF [43] and DIF-Net [18], although the details are reconstructed better than others, there are still a few streaking artifacts and unclear contours. The reconstructed results of C^{2}RV have clearer shape contours, better internal details, and almost no streaking artifacts. Furthermore, Figure 6 shows the visualization of results reconstructed from different numbers of projection views, demonstrating a consistent conclusion with the above.

Downstream Evaluation. In addition to quantitative and qualitative evaluation, we validate the reconstructed CT on the downstream task, *i.e.*, segmentation. Specifically, we utilize LungMask toolkit [8] to conduct left/right-lung segmentation on CT reconstructed by different methods. As the results are shown in Table 2 and Figure 7, compared with other methods, the segmentation masks on the reconstructed CT of C^{2}RV are more consistent with the segmentation on the ground-truth CT. This means our proposed C^{2}RV has the potential to reconstruct high-quality CT that can be further applied in downstream scenarios.

5. Ablation Study

Ablation studies are conducted to explore the effectiveness of the proposed MS-3DV and SVC-Att, and different designs for MS-3DV. Moreover, we further analyze the robustness of our C^{2}RV to varying viewing angles and noisy scanning parameters. All the following ablative experiments are conducted on 6-view reconstruction of chest CT with the resolution of 256^3.

5.1. Proposed MS-3DV and SVC-Att

Ablation on MS-3DV and SVC-Att. We regard DIF-Net [18] as the baseline model and compare the reconstruction performance of introducing MS-3DV and SVC-Att. In DIF-Net, multi-view features are aggregated (σ in Eqn. 2) with MLPs or Max-Pooling + MLPs. Comparison is shown in Table 3. In (+MS-3DV), multi-scale voxel-aligned features are concatenated with max-pooled multi-scale features. In (+SVC), we randomly initialize a learnable vector before training, as an alternative to the reference feature (*i.e.*, $\hat{\mathcal{F}}(p)$ in Eqn. 9); also see Figure 4. Both MS-3DV and SVC-Att can improve the reconstruction performance, and the framework achieves new state-of-the-art performance by jointly incorporating the above two.

Different Designs for MS-3DV. As shown in Table 4, we compare the performance of using different numbers of scales, and selections of initial feature map F^1 and resolution r^1. It is important to incorporate multi-scale features, which provide richer information than single-scale for identifying different anatomies, such as organs (*e.g.*, lung) and bones (*e.g.*, spine). We do not further increase the number of scales (*e.g.*, 4) since the size of the feature map at the third scale is too small (*i.e.*, 4×4). For the choice of F^1, the output of the encoder is better as it contains more high-level features than the decoder. Empirically, the initial resolution of 16 is the best choice for the trade-off between the global (high-level) and local (details) features.

5.2. Robustness Analysis

Let $\mathcal{A} = \{\alpha_1, \ldots, \alpha_N\}$ denote the viewing angles in the origional evaluation. The first experiment is conducted by choosing different viewing angles, *i.e.*, $\mathcal{A}' = \{\alpha_i + \Delta\alpha \mid \alpha_i \in \mathcal{A}\}$, where $\Delta\alpha$ is the angle offset. As shown in Table 5, the performance of C^{2}RV is stable with varying angles. The second study is about the noisy scanning parameters. Taking the viewing angles as an example, we assume the measurement process is noisy, which means that multi-view projections are measured from $\mathcal{A}' = \{\alpha_i + \eta_i \mid \alpha_i \in \mathcal{A}\}$, where η_i is the noise that obeys the uniform distribution $U(-\epsilon, +\epsilon)$. In this case, the projection function π is still defined based on original viewing angles, *i.e.*, $\mathcal{A}$, since the noise is unobservable. In Table 5, we consider two scanning parameters, including the viewing angle, and the distance of source to origin, which are major factors related to the formulation of the projection function (see Appendix in [18]). Experiments show that our C^{2}RV is robust to slight shifts in scanning parameters.

Table 3. Ablation study on different aggregation methods (M.: MLPs [18], Max-M.: Max-Pooling + MLPs [18], SVC: our proposed scale-view cross-attention) and multi-scale 3D volumetric representations (MS-3DV). PSNR (dB) and SSIM (×10^{-2}) are evaluated on 6-view reconstruction of chest CT.

Method	Aggregation			MS-3DV	PSNR	SSIM
	M.	Max-M.	SVC			
DIF-Net [18]	✓				25.55	84.42
		✓			25.62	84.40
+MS-3DV		✓		✓	26.62	87.33
+SVC			✓		27.84	90.22
C^{2}RV (*ours*)			✓	✓	**29.23**	**92.78**

Table 4. Ablation study on the number of scales, the initial feature map F^1, and the initial resolution r^1. The selection of F^1 can be the final-layer feature map of the encoder or decoder. PSNR and SSIM are evaluated on 6-view reconstruction of chest CT.

# Scales	F^1	r^1	PSNR (dB)	SSIM (10^{-2})
1	Encoder	16	28.98 (−0.25)	92.38 (−0.40)
2	Encoder	16	29.09 (−0.14)	92.57 (−0.21)
3	**Decoder**	16	28.57 (−0.66)	91.85 (−0.93)
3	Encoder	**12**	28.96 (−0.27)	92.72 (−0.06)
3	Encoder	**24**	29.23 (−0.00)	92.75 (−0.03)
3	Encoder	16	**29.23**	**92.78**

Table 5. Robustness analysis on varying angles and noisy scanning parameters, including viewing angles and the distance of source to origin (DSO). For noisy scanning parameters, the noisy offsets obey the uniform distribution, *i.e.*, $U(-\epsilon, +\epsilon)$. PSNR and SSIM are evaluated on 6-view reconstruction of chest CT.

Varying Angles	Noisy Parameters		PSNR (dB)	SSIM (10^{-2})
	Angles	DSO		
0°	-	-	29.23	92.78
+10°	-		29.24 (+0.01)	92.80 (+0.02)
+20°		-	29.23 (−0.00)	92.79 (+0.01)
0°	**±0.5°**	-	28.98 (−0.25)	92.57 (−0.21)
	±1.0°		28.18 (−1.05)	91.88 (−0.90)
0°	-	**±2mm**	29.04 (−0.19)	92.64 (−0.14)
		±3mm	27.85 (−1.38)	91.61 (−1.17)

6. Conclusion

In this work, we propose a novel framework, namely C^{2}RV, for sparse-view cone-beam CT reconstruction. The novelties are mainly composed of 1.) multi-scale 3D volumetric representations (MS-3DV) to enable efficient cross-regional feature learning in the 3D space, and 2.) scale-view cross-attention (SVC-Att) to adaptively aggregate multi-scale and multi-view features. Our C^{2}RV shows superior reconstruction performance compared with previous state-of-the-art, the practical potential of reconstructed CT in downstream applications, and robustness to slightly noisy measurement processes. Although our C^{2}RV performs well in a specific dataset, it will fail when adapting to other datasets with unseen anatomies (*e.g.*, chest→head) as C^{2}RV only learns the dataset-specific distribution priors. Hence, it would also be important to improve the few-shot or even zero-shot adaptation ability by introducing new training schemes or network frameworks, which will be left as our future works.

Acknowledgements. This work is partially supported by a research grant from the National Natural Science Foundation of China under Grant 62306254 and a grant from the Hong Kong Innovation and Technology Fund under Grant ITS/030/21.

References

[1] Anders H Andersen and Avinash C Kak. Simultaneous algebraic reconstruction technique (sart): a superior implementation of the art algorithm. *Ultrasonic imaging*, 6(1):81–94, 1984. 3, 6

[2] Hyungjin Chung, Dohoon Ryu, Michael T McCann, Marc L Klasky, and Jong Chul Ye. Solving 3d inverse problems using pre-trained 2d diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22542–22551, 2023. 2

[3] Yu Fang, Lanzhuju Mei, Changjian Li, Yuan Liu, Wenping Wang, Zhiming Cui, and Dinggang Shen. Snaf: Sparse-view cbct reconstruction with neural attenuation fields. *arXiv preprint arXiv:2211.17048*, 2022. 2, 3

[4] Lee A Feldkamp, Lloyd C Davis, and James W Kress. Practical cone-beam algorithm. *Josa a*, 1(6):612–619, 1984. 3, 6, 7

[5] Richard Gordon, Robert Bender, and Gabor T Herman. Algebraic reconstruction techniques (art) for three-dimensional electron microscopy and x-ray photography. *Journal of theoretical Biology*, 29(3):471–481, 1970. 3

[6] Yo Seob Han, Jaejun Yoo, and Jong Chul Ye. Deep residual learning for compressed sensing ct reconstruction via persistent homology analysis. *arXiv preprint arXiv:1611.06391*, 2016. 1, 2

[7] Ji He, Yongbo Wang, and Jianhua Ma. Radon inversion via deep learning. *IEEE transactions on medical imaging*, 39 (6):2076–2087, 2020. 1, 2

[8] Johannes Hofmanninger, Forian Prayer, Jeanny Pan, Sebastian Röhrich, Helmut Prosch, and Georg Langs. Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. *European Radiology Experimental*, 4(1):1–13, 2020. 7

[9] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 2

[10] Xia Huang, Jian Wang, Fan Tang, Tao Zhong, and Yu Zhang. Metal artifact reduction on cervical ct images by deep residual learning. *Biomedical engineering online*, 17:1–15, 2018. 1, 2

[11] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction of clothed humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3093–3102, 2020. 1, 3

[12] Yixiang Jiang. Mfct-gan: multi-information network to reconstruct ct volumes for security screening. *Journal of Intelligent Manufacturing and Special Equipment*, 2022. 3

[13] Kyong Hwan Jin, Michael T McCann, Emmanuel Froustey, and Michael Unser. Deep convolutional neural network for inverse problems in imaging. *IEEE Transactions on Image Processing*, 26(9):4509–4522, 2017. 1, 2, 6, 7

[14] Daeun Kyung, Kyungmin Jo, Jaegul Choo, Joonseok Lee, and Edward Choi. Perspective projection-based 3d ct reconstruction from biplanar x-rays. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 3

[15] Anish Lahiri, Marc Klasky, Jeffrey A Fessler, and Saiprasad Ravishankar. Sparse-view cone beam ct reconstruction using data-consistent supervised and adversarial learning from scarce training data. *arXiv preprint arXiv:2201.09318*, 2022. 2

[16] Bo Li, Kaitao Xue, Bin Liu, and Yu-Kun Lai. Bbdm: Image-to-image translation with brownian bridge diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1952–1961, 2023. 6

[17] Wei-An Lin, Haofu Liao, Cheng Peng, Xiaohang Sun, Jingdan Zhang, Jiebo Luo, Rama Chellappa, and Shaohua Kevin Zhou. Dudonet: Dual domain network for ct metal artifact reduction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10512–10521, 2019. 2

[18] Yiqun Lin, Zhongjin Luo, Wei Zhao, and Xiaomeng Li. Learning deep intensity field for extremely sparse-view cbct reconstruction. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 13–23, Cham, 2023. Springer Nature Switzerland. 1, 2, 3, 5, 6, 7, 8

[19] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6), 2015. 1, 3

[20] Chenglong Ma, Zilong Li, Junping Zhang, Yi Zhang, and Hongming Shan. Freeseed: Frequency-band-aware and self-guided network for sparse-view ct reconstruction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 250–259. Springer, 2023. 1, 2, 6, 7

[21] 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. *Communications of the ACM*, 65(1):99–106, 2021. 3

[22] Jinxiao Pan, Tie Zhou, Yan Han, and Ming Jiang. Variable weighted ordered subset image reconstruction algorithm. *International Journal of Biomedical Imaging*, 2006, 2006. 3

[23] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 3

[24] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019. 1, 3

[25] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 2, 5

[26] Darius Rückert, Yuanhao Wang, Rui Li, Ramzi Idoughi, and Wolfgang Heidrich. Neat: Neural adaptive tomography. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022. 3

[27] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2304–2314, 2019. 3

[28] William C Scarfe, Allan G Farman, Predag Sukovic, et al. Clinical applications of cone-beam computed tomography in dental practice. *Journal-Canadian Dental Association*, 72 (1):75, 2006. 1

[29] Arnaud Arindra Adiyoso Setio, Alberto Traverso, Thomas De Bel, Moira SN Berens, Cas Van Den Bogaard, Piergiorgio Cerello, Hao Chen, Qi Dou, Maria Evelina Fantacci, Bram Geurts, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge. *Medical image analysis*, 42:1–13, 2017. 5, 6

[30] 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*, 3(11):880–888, 2019. 3

[31] Liyue Shen, John Pauly, and Lei Xing. Nerp: implicit neural representation learning with prior embedding for sparsely sampled image reconstruction. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. 2, 3, 6, 7

[32] Yang Song, Liyue Shen, Lei Xing, and Stefano Ermon. Solving inverse problems in medical imaging with score-based generative models. *arXiv preprint arXiv:2111.08005*, 2021. 2

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[34] Ce Wang, Kun Shang, Haimiao Zhang, Qian Li, Yuan Hui, and S Kevin Zhou. Dudotrans: dual-domain transformer provides more attention for sinogram restoration in sparse-view ct reconstruction. *arXiv preprint arXiv:2111.10790*, 2021. 2

[35] Jianing Wang, Yiyuan Zhao, Jack H Noble, and Benoit M Dawant. Conditional generative adversarial networks for metal artifact reduction in ct images of the ear. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*, pages 3–11. Springer, 2018. 1, 2

[36] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[37] 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*, 40(11):3002–3014, 2021. 1, 2

[38] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. Icon: Implicit clothed humans obtained from normals. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13286–13296. IEEE, 2022. 1, 3

[39] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J Black. Econ: Explicit clothed humans optimized via normal integration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 512–523, 2023. 1, 3

[40] Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. Point-nerf: Point-based neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5438–5448, 2022. 1, 3

[41] Fukun Yin, Wen Liu, Zilong Huang, Pei Cheng, Tao Chen, and Gang Yu. Coordinates are not lonely-codebook prior helps implicit neural 3d representations. *Advances in Neural Information Processing Systems*, 35:12705–12717, 2022. 1, 3

[42] Xingde Ying, Heng Guo, Kai Ma, Jian Wu, Zhengxin Weng, and Yefeng Zheng. X2ct-gan: reconstructing ct from biplanar x-rays with generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10619–10628, 2019. 3

[43] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021. 3, 6, 7

[44] Ruyi Zha, Yanhao Zhang, and Hongdong Li. Naf: Neural attenuation fields for sparse-view cbct reconstruction. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VI*, pages 442–452. Springer, 2022. 2, 3, 6

[45] 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*, 37(6):1407–1417, 2018. 1, 2

[46] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE transactions on pattern analysis and machine intelligence*, 44(6): 3170–3184, 2021. 1, 3