VolumeNeRF: CT Volume Reconstruction from a Single Projection View

Jiachen Liu[1] and Xiangzhi Bai[1,2,3]

[1] Image Processing Center, Beihang University, Beijing, China
[2] State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China
[3] Advanced Innovation Center for Biomedical Engineering, Beihang University, Beijing, China
jackybxz@buaa.edu.cn

Abstract. Computed tomography (CT) plays a significant role in clinical practice by providing detailed three-dimensional information, aiding in accurate assessment of various diseases. However, CT imaging requires a large number of X-ray projections from different angles and exposes patients to high doses of radiation. Here we propose VolumeNeRF, based on neural radiance fields (NeRF), for reconstructing CT volumes from a single-view X-ray. During training, our network learns to generate a continuous representation of the CT scan conditioned on the input X-ray image and render an X-ray image similar to the input from the same viewpoint as the input. Considering the ill-posedness and the complexity of the single-perspective generation task, we introduce likelihood images and the average CT images to incorporate prior anatomical knowledge. A novel projection attention module is designed to help the model learn the spatial correspondence between voxels in CT images and pixels in X-ray images during the imaging process. Extensive experiments conducted on a publicly available chest CT dataset show that our VolumeNeRF achieves better performance than other state-of-the-art methods. Our code is available at https://www.github.com/Aurora132/VolumeNeRF.

Keywords: Computed tomography reconstruction · X-ray image · Neural radiance fields · Anatomical priors · Projection attention.

1 Introduction

Computed Tomography (CT) is a popular medical imaging technique that utilizes X-ray technology to produce cross-sectional images of bodies, allowing for precise visualization of tissues. However, volumetric imaging typically requires numerous X-ray projection views from different positions, resulting in long imaging time and excessive radiation exposure to patients [14,19]. Additionally, due to their high complexity and cost, CT scanners are not widely available, especially in less-developed regions [17]. In order to address these issues, attempts to reconstruct CT volumes from 2D X-ray images have been made.

X-ray imaging projects all tissues onto a 2D plane, enabling the visualization of internal structures within bodies. This imaging technique is relatively low-cost, widely available, and exposes patients to minimal radiation doses [18]. Recently, several methods that leverage deep learning for 3D CT volume reconstruction from 2D X-ray images have emerged [9,21,24,20]. The main challenge of CT volume reconstruction from X-rays is the absence of depth information, rendering it an ill-posed problem [2]. To tackle the challenge, these approaches design 2D-to-3D network structures and train them on large-scale data to restore missing depth information. Nevertheless, these models implicitly learned the transformation mapping from 2D to 3D without considering the spatial correspondence between pixels in X-ray images and voxels in CT volumes. Consequently, their performance in solving this ill-posed problem is decreased because of the failure to introduce prior projection relationships during the imaging process.

The neural radiance fields (NeRF) [15] model is a mainstream method for synthesizing novel views of complex scenes. NeRF represents a scene as a continuous 3D volume and employs a neural network to model both the geometry and appearance of the scene. Then NeRF uses volume rendering to integrate information stored in the 3D volume along each viewing ray for new view synthesis. Considering the similarity between the natural light imaging process and the X-ray imaging process, migrating the NeRF model to CT reconstruction problems appears feasible. However, generating such representations generally requires multiple images from various viewpoints [16,4,8,22]. Some recent works further study single-view NeRF models, but they mainly focus on novel-view synthesis or surface reconstruction [3]. Considering the complexity of human anatomical structures and the demand to reconstruct 3D information inside bodies, it is challenging to directly apply NeRF to single-view CT reconstruction. Additional constraints and prior knowledge are needed for single-view CT reconstruction.

Considering the issues discussed above, we present VolumeNeRF, a model that adopts NeRF to reconstruct 3D CT volumes from a single projection view. Specifically, we design a new architecture to recover the lost depth information from the 2D X-ray images and to generate a continuous attenuation coefficient distribution of the corresponding scene. Then we employ volume rendering based on the Lambert-Beer law and output a scene view with the same view direction as the input image. This process allows for extra constraints by narrowing the gap between the rendered and the input X-ray images. To fully utilize the similarities of anatomical structures across different bodies, we calculate the likelihood of the intensity of each pixel in the input X-ray image to quantify the discrepancies between each individual and the group average. Subsequently, we input the likelihood images and the average CT images into the network to incorporate prior anatomical knowledge. Additionally, projection attention modules are proposed to integrate prior projection relationships into the model and to enhance the ability of the model in learning the spatial correspondence between voxels from CT volumes and pixels from X-ray images. We evaluate our method on a publicly available chest CT dataset. Qualitative and quantitative analyses are conducted, and the results demonstrate the superiority of VolumeNeRF.

2 Related Works

Single-view NeRF aims to produce a 3D scene representation conditioned on one image. These methods introduce auxiliary information or self-supervision techniques to learn relationships across scenes and focus on novel view synthesis, depth estimation, and surface reconstruction [25,23,3,6]. Therefore, they are limited in 3D volume reconstruction since CT reconstruction lacks side information and needs to reconstruct information inside human bodies.

CT Reconstruction from X-ray has been recently studied using deep learning [9,21,24,20,12,10,5]. These works design and train models to convert uniplanar or biplanar X-ray images into 3D CT volumes. However, they do not incorporate projection or anatomical structure prior knowledge and thus lack extra constraints during model training, resulting in a decline in model performance.

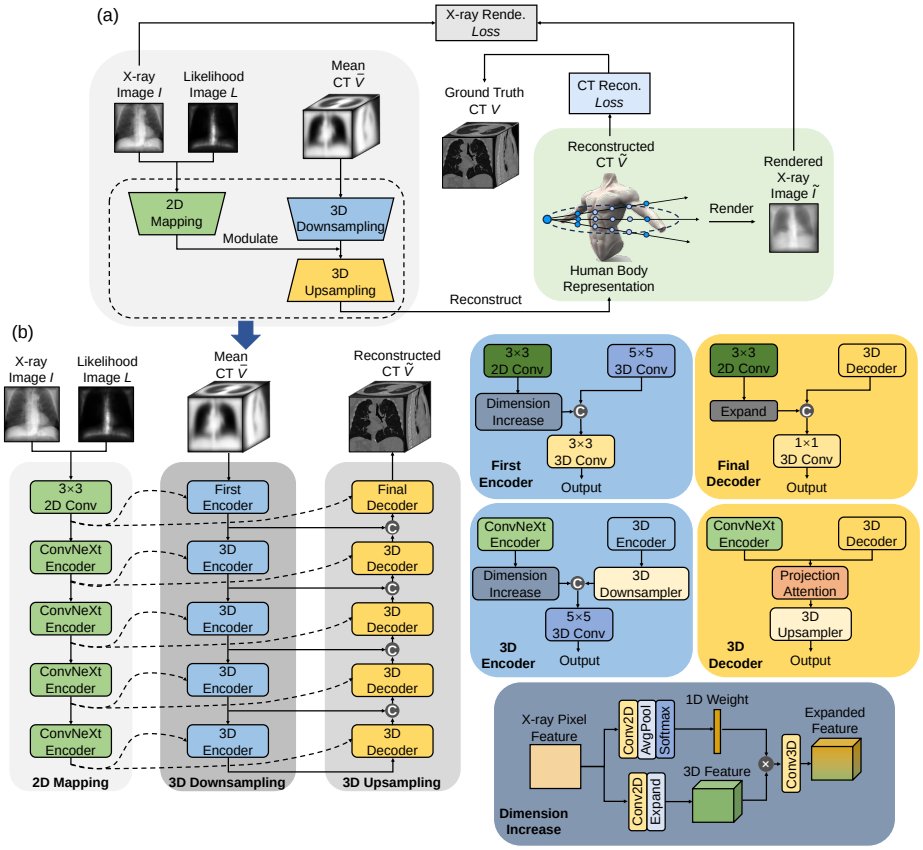

Fig. 1. (a) Overview of VolumeNeRF; (b) Illustration of the model structure.

3 Method

When given an X-ray image I, the goal of VolumeNeRF is to reconstruct the corresponding CT image $\tilde{V}$. We adopt a 3D encoder-decoder network to reconstruct volumetric representations. Similar to StyleGAN2 [11], we utilize ConvNeXt [13] as a 2D mapping network to condition and modulate the 3D CT reconstruction. The generated style vectors are then fed into the 3D network to form final results. We also incorporate the X-ray likelihood image L and the mean CT image $\bar{V}$ into the model to integrate prior anatomical information. After the reconstruction process, we obtain a 3D representation of attenuation characteristics, which can be directly considered as the CT volume. Subsequently, we employ volumetric rendering to synthesize an X-ray image from the same viewpoint as the input, thus providing additional supervision signals.

3.1 Prior Anatomical Knowledge Incorporation

The anatomical structures of different human bodies typically exhibit similarities, such as the number, shape, and position of organs. Leveraging this prior knowledge can provide auxiliary information to enhance CT reconstruction. For this purpose, we perform deformable registration to align all CT volumes and calculate the average value of all CT images in the training set, which represents the overall distribution of materials within the human body. The mean CT image serves as the basis for CT reconstruction and is inputted into the 3D encoder-decoder network. In addition, we compute a likelihood image L to quantify the relationship between each individual and the population. Specifically, we assume that the attenuation coefficient distribution of each pixel p in the X-ray image follows a one-dimensional Gaussian distribution $N\left(\mu_p, \sigma_p^2\right)$. To compute the parameters, mean μ_p and variance σ_p^2, we apply maximum likelihood estimation to fit the training set data to the corresponding distribution. We utilize the negative log-likelihood function to measure the deviation of each pixel between individuals and the population, which can be expressed as:

$$L_p = -\log P\left(x_p \mid \mu_p, \sigma_p\right) = \log \sqrt{2\pi\sigma_p^2} + \frac{\left(x_p - \mu_p\right)^2}{2\sigma_p^2}. \tag{1}$$

A lower L_p value indicates a smaller deviation of pixel p between the current input individual and the population. Therefore, voxels situated on the ray connecting the X-ray source and pixel p are expected to approximate the corresponding average value, which can be obtained from the input mean CT image.

3.2 3D Representation Generation

Downsampling Encoder and Upsampling Decoder. We design cascaded encoders to extract multi-level information from the average CT image and learn the coordinate projection relationship of CT reconstruction. For the i-th encoder, the dimension increase block takes the style tensor w^i as input and converts it

into 3D. Specifically, w^i is sent to a convolutional layer and expanded in the depth direction. Then we use a set of weights to characterize the relationship between the expanded features in this direction. The 3D feature v^{i-1} from the previous encoder is downsampled using the down-sampling block. The outputs of these two blocks are concatenated and then inputted into the fusion block to generate new 3D feature v^i. The decoder is proposed to reconstruct 3D feature maps to the size of the input CT images. As illustrated in Fig. 1(b), each decoder takes the style tensor and the output from the previous decoder as input and conducts upsampling. Notably, we introduce a projection attention module in each decoder to enhance the feature of each voxel based on the projection relationship.

Projection Attention Module. In the X-ray radiography system, the value of pixel p is determined by the intensities of voxels located on the ray passing through it. Thus, we design the projection attention module to help the model learn the correspondence between voxels and pixels. In the i-th decoder, the module receives the 2D style tensor $w^i \in \mathbb{R}^{C\times H\times W}$ and the 3D feature $v^{i+1} \in \mathbb{R}^{C\times D\times H\times W}$ as input. First, for each voxel in the v^{i+1}, we calculate the position of the intersection point between the detector plane and the ray passing through the voxel and the X-ray source. With these results, we employ bilinear interpolation to extract the pixel feature corresponding to each voxel from w^i and stack them based on the position number of voxels to obtain $w_v^i \in \mathbb{R}^{C\times D\times H\times W}$. To enhance the learning of the spatial correspondence by the model, we draw inspiration from DCN v2 [26] and introduce learnable offsets. Concretely, v^{i+1} and w_v^i are concatenated and subsequently sent to two 3D convolution layers to generate offsets $\triangle p \in \mathbb{R}^{2K\times D\times H\times W}$ and weights $\triangle m \in \mathbb{R}^{K\times D\times H\times W}$:

$$\triangle p = F_1\left(\left[v^{i+1}, w_v^i\right]\right). \tag{2}$$

$$\triangle m = \text{Softmax}\left(F_2\left(\left[v^{i+1}, w_v^i\right]\right)\right). \tag{3}$$

Here K represents the number of related pixels. For each voxel with the location v_0 in the 3D feature map v^{i+1}, its corresponding pixel feature $y\left(v_0\right)$ becomes

$$y\left(v_0\right) = \sum_{k=1}^{K} \triangle m_k \times w^i\left(w_0 + \triangle p_k\right). \tag{4}$$

Here $y \in \mathbb{R}^{C\times D\times H\times W}$ represents the final corresponding pixel feature map. w_0 denotes the position of the intersection point between the ray passing through v_0 and the detector plane. $\triangle p_k$ and $\triangle m_k$ are the offset and the weight of the k-th related pixel. Finally, we concatenate the original 3D feature v^{i+1} and the corresponding pixel feature map y and fuse them using a 3D convolutional operator, generating the enhanced voxel feature map $v_y^{i+1} \in \mathbb{R}^{C\times D\times H\times W}$.

3.3 Volume Rendering

To introduce additional supervision information, we further apply volumetric rendering to synthesize X-ray images $\widetilde{I}$ from the same viewpoint as the input I.

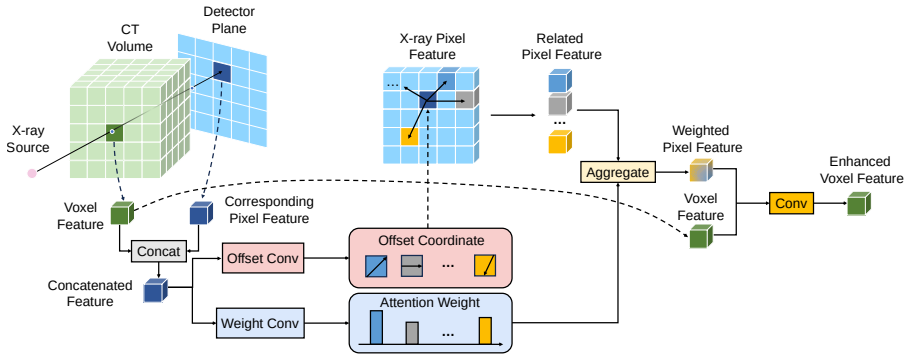

Fig. 2. Illustration of the projection attention module.

Following the Lambert-Beer law, the rendering equation can be written as

$$I(r) = I_0(r) \times \exp\left(-\sum_{i=1}^{N} \mu_i \delta_i\right). \tag{5}$$

Here $I_0(r)$ is the intensity of the incident beam r, and $I(r)$ denotes its intensity after traveling through the human body. μ_i represents the attenuation coefficient of the i-th voxel through which the beam r passes, while δ_i is the distance that the beam r propagates within this voxel.

3.4 Overall Objective

Our learning objective is written as follows:

$$L_{\text{total}} = \lambda_{\text{recon}} L_{\text{recon}} + \lambda_{\text{edge}} L_{\text{edge}} + \lambda_{\text{render}} L_{\text{render}}, \tag{6}$$

where L_{recon}, L_{edge}, and L_{render} represent the L1 loss between the reconstructed CT and the ground truth, the edges of the generated and real CT images (extracted using the Scharr operator), and the input X-ray image and the rendered image, respectively. λ_{recon}, λ_{edge}, and λ_{render} balance these terms, with values set to 1, 0.05, and 0.001, respectively.

4 Experiments and Results

4.1 Data and Settings

Considering the difficulty of collecting paired X-ray images and corresponding CT scans, we employ the digitally reconstructed radiographs (DRR) technology to produce X-ray projections following previous studies [21,24,20,12,10]. We validate our model on the open source LIDC-IDRI dataset that comprises 1,018 chest CT volumes [1]. We first resample all scans to a voxel size of $2.5 \times 2.5 \times 2.5$

mm. Due to memory limitations, we then crop an area of $128 \times 128 \times 128$ voxels from the center of each scan. Subsequently, the differentiable DRR [7] is adopted to synthesize the X-ray image with a resolution of 128×128. We randomly divide the dataset into a training set (865 volumes), a validation set (51 volumes), and a testing set (102 volumes). The peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) indices are utilized for the assessment of our method.

4.2 Results

Comparison results. To demonstrate the advantages of our method, we compare our VolumeNeRF with four deep learning-based single-view CT reconstruction methods: 2DCNN [9], PatRecon [21], X2CT-CNN [24], and sci-f [10].

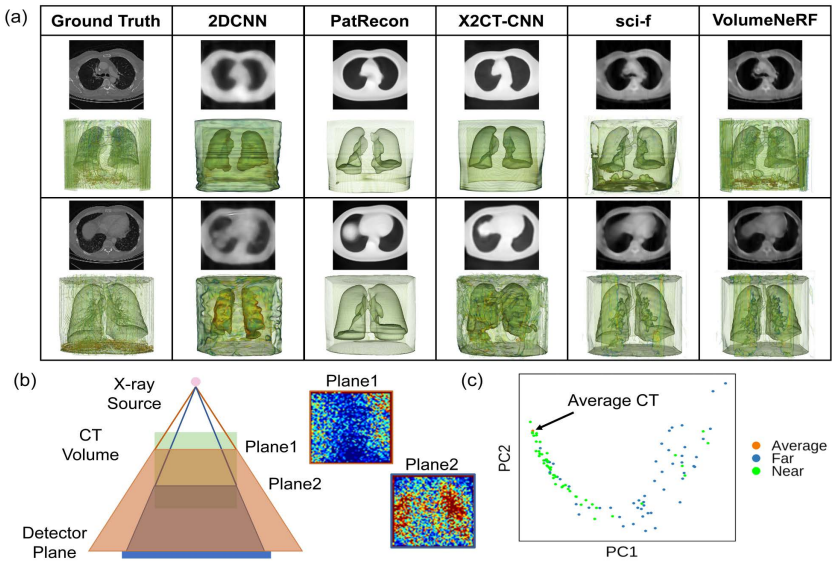

Fig. 3. (a) Results from various approaches; (b) Visualization of offset coordinates of two planes; (c) PCA visualization of the feature representations of testing samples.

Initially, we conduct a qualitative assessment of the reconstruction performance of various methods. As shown in Fig. 3(a), our method exhibits a distinct advantage in preserving edges and small anatomical structures. The first three approaches (2DCNN, PatRecon, and X2CT-CNN) lose structural and textural details due to the absence of prior knowledge and additional constraints. The sci-f model leverages spatial projection relationships and retains many reconstruction details. However, it blurs boundaries and intricate details. The subsequent quantitative results are displayed in Table 1. Our method achieves the superior performance on the PSNR and SSIM indices. Notably, there is a 12.4%

enhancement over the second-best approach (sci-f) in SSIM, indicating that the reconstructed CT images from our method exhibit better visual quality.

Ablation Studies. To evaluate the impacts of likelihood images (LIs), projection attention modules (PAMs), and the render loss term (RL) in our model, we perform ablation studies, with results presented in supplementary materials and Table 2. The reconstruction performance decreases when removing any of these three components. We also find incorporating likelihood images leads to the most noticeable improvement. These results show the effectiveness of integrating prior anatomical knowledge and projection relationships into the model.

Table 1. Quantitative results, including averages and standard deviations.

Method	PSNR	SSIM
2DCNN	22.42(0.24)	0.433(0.006)
PatRecon	22.75(0.21)	0.478(0.007)
X2CT-CNN	23.01(0.15)	0.542(0.005)
sci-f	24.18(0.32)	0.587(0.004)
VolumeNeRF	25.59(0.20)	0.660(0.002)

Table 2. Ablation study results, including means and standard deviations.

Method	PSNR	SSIM
w/o RL	25.02(0.12)	0.607(0.004)
w/o LIs	23.87(0.26)	0.584(0.008)
w/o PAMs	24.79(0.18)	0.599(0.003)
VolumeNeRF	25.59(0.20)	0.660(0.002)

Coordinate offsets visualization. We obtain the learned offset coordinates for two randomly chosen planes of a randomly selected testing sample and count the number of occurrences of corresponding pixels for each voxel in these planes separately. We visualize the results in Fig. 3(b) and observe that the offset coordinates in Plane 1 are distributed on both sides of the image, while those in Plane 2 are widely distributed in the image. This conforms to the spatial projection relationship because Plane 1 is farther from the detector plane, while Plane 2 is closer. Intersection points between the detector plane and the rays which pass through the X-ray source and voxels in the distant plane are more likely to fall outside the detection range. The corresponding pixels for such voxels are located at the edges of the X-ray image, since we manually constrain the learned offset coordinates within the X-ray image size range. For voxels in the near plane, their corresponding pixels are generally situated within this range.

Analysis of likelihood images. Based on our previous analysis, if a pixel deviates from the prior distribution, its log-likelihood value increases, indicating substantial differences between the voxels (situated on the ray connecting the X-ray source and the pixel) of the patient and the average. To assess the impact of likelihood images, we extract 3D feature maps from the final layer of the 3D encoder. These maps contain high-level semantic information of volumes and are visualized using the principal component analysis (PCA) projection after

flattening. We categorize all testing samples into two classes based on the average of their corresponding likelihood images. Fig. 3(c) demonstrates that volumes with a low mean of the corresponding likelihood images (near class) are adjacent to the average volume, whereas those with a high mean (far class) are distant from it, as expected. The results show the effectiveness of likelihood images in quantifying differences between each volume and the average volume.

5 Conclusion

In this paper, we present a NeRF-based model for 3D CT reconstruction from a single-view X-ray image. To address this ill-posed problem, we compute likelihood images and input them, along with the average CT images, into the network to fully leverage the similarity of anatomical structures in human bodies. Furthermore, we exploit projection attention modules to introduce prior projection relationships. These modules can automatically learn spatial correlations between voxels from CT volumes and pixels from X-ray images. The experiments show that our method achieves superior performance compared to others. In the future, we plan to explore the clinical value of our method in various tasks, such as spinal deformity classification and orthopedic preoperative evaluation.

Acknowledgments. We thank Yuxuan Liu for his technical contributions, including training networks and conducting contrast experiments. This work was supported by the National Natural Science Foundation of China, the Beijing Natural Science Foundation, and the Fundamental Research Funds for the Central Universities.

Disclosure of Interests. The authors declare no competing interests.

References

1. Armato III, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., Zhao, B., Aberle, D.R., Henschke, C.I., Hoffman, E.A., et al.: The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. Medical Physics **38**(2), 915–931 (2011)
2. Bertero, M., Poggio, T.A., Torre, V.: Ill-posed problems in early vision. Proceedings of the IEEE **76**(8), 869–889 (1988)
3. Cao, A.Q., de Charette, R.: Scenerf: Self-supervised monocular 3d scene reconstruction with radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9387–9398 (2023)
4. Chen, A., Xu, Z., Zhao, F., Zhang, X., Xiang, F., Yu, J., Su, H.: Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14124–14133 (2021)
5. Corona-Figueroa, A., Frawley, J., Bond-Taylor, S., Bethapudi, S., Shum, H.P., Willcocks, C.G.: Mednerf: Medical neural radiance fields for reconstructing 3d-aware ct-projections from a single x-ray. In: 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). pp. 3843–3848. IEEE (2022)

6. Deng, C., Jiang, C., Qi, C.R., Yan, X., Zhou, Y., Guibas, L., Anguelov, D., et al.: Nerdi: Single-view nerf synthesis with language-guided diffusion as general image priors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20637–20647 (2023)
7. Gopalakrishnan, V., Golland, P.: Fast auto-differentiable digitally reconstructed radiographs for solving inverse problems in intraoperative imaging. In: Workshop on Clinical Image-Based Procedures. pp. 1–11. Springer (2022)
8. Guo, H., Peng, S., Lin, H., Wang, Q., Zhang, G., Bao, H., Zhou, X.: Neural 3d scene reconstruction with the manhattan-world assumption. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5511–5520 (2022)
9. Henzler, P., Rasche, V., Ropinski, T., Ritschel, T.: Single-image tomography: 3d volumes from 2d cranial x-rays. In: Computer Graphics Forum. vol. 37, pp. 377–388. Wiley Online Library (2018)
10. Jiang, Y., Yuan, X., Pei, Y.: Spatially-consistent implicit volumetric function for uni-and bi-planar x-ray-based computed tomography reconstruction. In: 2023 IEEE 20th International Symposium on Biomedical Imaging. pp. 1–5. IEEE (2023)
11. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8110–8119 (2020)
12. Kyung, D., Jo, K., Choo, J., Lee, J., Choi, E.: Perspective projection-based 3d ct reconstruction from biplanar x-rays. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 1–5. IEEE (2023)
13. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11976–11986 (2022)
14. Lo, P., Van Ginneken, B., Reinhardt, J.M., Yavarna, T., De Jong, P.A., Irving, B., Fetita, C., Ortner, M., Pinho, R., Sijbers, J., et al.: Extraction of airways from ct (exact'09). IEEE Transactions on Medical Imaging **31**(11), 2093–2107 (2012)
15. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
16. Oechsle, M., Peng, S., Geiger, A.: Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5589–5599 (2021)
17. Organization, W.H., et al.: Baseline country survey on medical devices. Geneva: World Health Organization (2010)
18. Ou, X., Chen, X., Xu, X., Xie, L., Chen, X., Hong, Z., Bai, H., Liu, X., Chen, Q., Li, L., et al.: Recent development in x-ray imaging technology: Future and challenges. Research (2021)
19. Power, S.P., Moloney, F., Twomey, M., James, K., O'Connor, O.J., Maher, M.M.: Computed tomography and patient risk: Facts, perceptions and uncertainties. World Journal of Radiology **8**(12), 902 (2016)
20. Ratul, M.A.R., Yuan, K., Lee, W.: Ccx-raynet: a class conditioned convolutional neural network for biplanar x-rays to ct volume. In: 2021 IEEE 18th International Symposium on Biomedical Imaging. pp. 1655–1659. IEEE (2021)
21. Shen, L., Zhao, W., Xing, L.: Patient-specific reconstruction of volumetric computed tomography images from a single projection view via deep learning. Nature Biomedical Engineering **3**(11), 880–888 (2019)

22. Xu, C., Wu, B., Hou, J., Tsai, S., Li, R., Wang, J., Zhan, W., He, Z., Vajda, P., Keutzer, K., et al.: Nerf-det: Learning geometry-aware volumetric representation for multi-view 3d object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23320–23330 (2023)
23. Yang, W., Chen, G., Chen, C., Chen, Z., Wong, K.Y.K.: S^3-nerf: Neural reflectance field from shading and shadow under a single viewpoint. Advances in Neural Information Processing Systems **35**, 1568–1582 (2022)
24. Ying, X., Guo, H., Ma, K., Wu, J., Weng, Z., Zheng, Y.: 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. pp. 10619–10628 (2019)
25. Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelnerf: Neural radiance fields from one or few images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4578–4587 (2021)
26. Zhu, X., Hu, H., Lin, S., Dai, J.: Deformable convnets v2: More deformable, better results. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9308–9316 (2019)