NAF: Neural Attenuation Fields for Sparse-View CBCT Reconstruction

Ruyi Zha(✉), Yanhao Zhang, and Hongdong Li

Australian National University, Canberra, Australia
ruyi.zha@anu.edu.au

Abstract. This paper proposes a novel and fast self-supervised solution for sparse-view CBCT reconstruction (Cone Beam Computed Tomography) that requires no external training data. Specifically, the desired attenuation coefficients are represented as a continuous function of 3D spatial coordinates, parameterized by a fully-connected deep neural network. We synthesize projections discretely and train the network by minimizing the error between real and synthesized projections. A learning-based encoder entailing hash coding is adopted to help the network capture high-frequency details. This encoder outperforms the commonly used frequency-domain encoder in terms of having higher performance and efficiency, because it exploits the smoothness and sparsity of human organs. Experiments have been conducted on both human organ and phantom datasets. The proposed method achieves state-of-the-art accuracy and spends reasonably short computation time.

Keywords: CBCT · Sparse view · Implicit neural representation

1 Introduction

Cone Beam Computed Tomography (CBCT) is an emerging medical imaging technique to examine the internal structure of a subject noninvasively. A CBCT scanner emits cone-shaped X-ray beams and captures 2D projections at equal angular intervals. Compared with the conventional Fan Beam CT (FBCT), CBCT enjoys the benefits of high spatial resolution and fast scanning speed [19]. Recent years have witnessed the blossoming of low dose CT, which delivers a significantly lower radiation dose during the scanning process. There are two ways to reduce the dose: decreasing source intensity or projection views [8]. This paper focuses on the latter, *i.e.*, sparse-view CBCT reconstruction.

Sparse-view CBCT reconstruction aims to retrieve a volumetric attenuation coefficient field from dozens of projections. It is a challenging task in two respects. First, insufficient views lead to notable artifacts. As a comparison, the traditional CBCT obtains hundreds of images. The inputs of sparse-view CBCT are 10×

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-16446-0_42.

L. Wang et al. (Eds.): MICCAI 2022, LNCS 13436, pp. 1–11, 2022.
https://doi.org/10.1007/978-3-031-16446-0_42

fewer. Second, the spatial and computational complexity of CBCT reconstruction is much higher than that of FBCT reconstruction due to the dimensional increase of inputs. CBCT relies on 2D projections to build a 3D model, while FBCT simplifies the process by stacking 2D slides restored from 1D projections (but in the sacrifice of time and dose).

Existing CBCT approaches can be divided into three categories: analytical, iterative and learning-based methods. Analytical methods estimate attenuation coefficients by solving the Radon transform and its inverse. A typical example is the FDK algorithm [7]. It produces good results in an ideal scenario but copes poorly with ill-posed problems such as sparse views. The second family, iterative methods, formulates reconstruction as a minimization process. These approaches utilize an optimization framework combined with regularization modules. While iterative methods perform well in ill-posed problems [2,20], they require substantial computation time and memory. Recently, learning-based methods have become popular with the rise of AI. They use deep neural networks to 1) predict and extrapolate projections [3,22,24,28], 2) regress attenuation coefficients with similar data [11,27], and 3) make optimization process differentiable [1,6,10]. Most of these methods [3,11,22,27] need extensive datasets for network training. Moreover, they rely on neural networks to remember what a CT looks like. Therefore it is difficult to apply a trained model of one application to another. While there are self-supervised methods [1,28], they operate under FBCT settings considering network capacity and memory consumption. Their performance and efficiency drop when applied to the CBCT scenario.

Apart from the aforementioned work designated for CT reconstruction, efforts have been made to deal with other ill-posed problems, such as 3D reconstruction in the computer vision field. Similar to CT reconstruction, 3D reconstruction uses RGB images to estimate 3D shapes, which are usually represented as discrete point clouds or meshes. Recent studies propose [13,16] Implicit Neural Representation (INR) as an alternative to those discrete representations. INR parameterizes a bounded scene as a neural network that maps spatial coordinates to metrics such as occupancy and color. With the help of position encoder [14,21], INR is capable to learn high-frequency details.

This paper proposes Neural Attenuation Fields (NAF), a fast self-supervised solution for sparse-view CBCT reconstruction. Here we use 'self-supervised' to highlight that NAF requires no external CT scans but the X-ray projections of the interested object. Inspired by 3D reconstruction work [13,16], we parameterize the attenuation coefficient field as an INR and imitates the X-ray attenuation process with a self-supervised network pipeline. Specifically, we train a Multi-Layer Perceptron (MLP), whose input is an encoded spatial coordinate (x, y, z) and whose output is the attenuation coefficient μ at that location. Instead of using a common frequency-domain encoding, we adopt hash encoding [14], a learning-based position encoder, to help the network quickly learn high-frequency details. Projections are synthesized by predicting the attenuation coefficients of sampled points along ray trajectories and attenuating incident beams accordingly. The network is optimized with gradient descent by minimizing the error

between real and synthesized projections. We demonstrate that NAF quantitatively and qualitatively outperforms existing solutions on both human organ and phantom datasets. While most INR approaches take hours for training, our method can reconstruct a detailed CT model within 10–40 minutes, which is comparable to iterative methods.

In summary, the main contributions of this work are:

- We propose a novel and fast self-supervised method for sparse-view CBCT reconstruction. Neither external datasets nor structural prior is needed except projections of a subject.
- The proposed method achieves state-of-the-art accuracy and spends relatively short computation time. The performance and efficiency of our method make it feasible for clinical CT applications.
- The code will be publicly available for investigation purposes.

2 Method

2.1 Pipeline

The pipeline of NAF is shown in Fig. 1. During a CBCT scanning, an X-ray source rotates around the object and emits cone-shaped X-ray beams. A 2D panel detects X-ray projections at equal angular intervals. NAF then uses the scanner geometry to imitate the attenuation process discretely. It learns CT shapes by comparing real and synthesized projections. After the model optimization, the final CT image is generated by querying corresponding voxels.

NAF consists of four modules: ray sampling, position encoding, attenuation coefficient prediction, and projection synthesis. First, we uniformly sample points

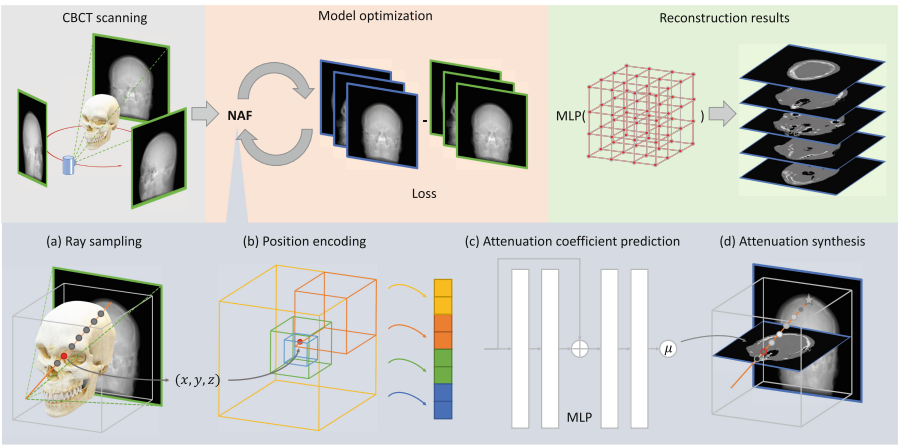

Fig. 1. NAF pipeline. Gray block: The CBCT scanner captures X-ray projections from different views. Blue block: NAF simulates projections. Orange block: NAF is optimized by comparing real and synthesized projections. Green block: NAF generates a CT model by querying corresponding voxels. (Color figure online)

along X-ray paths based on the scanner geometry. A position encoder network then encodes their spatial coordinates to extract valuable features. After that, an MLP network consumes the encoded information and predicts attenuation coefficients. The last step of NAF is to synthesize projections by attenuating incident X-rays according to the predicted attenuation coefficients on their paths.

2.2 Neural Attenuation Fields

Ray Sampling. Each pixel value of a projection image results from an X-ray passing through a cubical space and getting attenuated by the media inside. We sample N points at the parts where rays intersect the cube. A stratified sampling method [13] is adopted, where we divide a ray into N evenly spaced bins and uniformly sample one point at each bin. Setting N greater than the desired CT size ensures that at least one sample is assigned to every grid cell that an X-ray traverses. The coordinates of sampled points are then sent to the position encoding module.

Position Encoding. A simple MLP can theoretically approximate any function [9]. Recent studies [18,21], however, reveal that a neural network prefers to learn low-frequency details due to "spectral bias". To this end, a position encoder is introduced to map 3D spatial coordinates to a higher dimensional space.

A common choice is the *frequency encoder* proposed by Mildenhall *et al.* [13]. It decomposes a spatial coordinate $\mathbf{p} \in \mathbb{R}^3$ into L sets of sinusoidal components at different frequencies. While frequency encoder eases the difficulty of training networks, it is considered quite cumbersome. In medical imaging practise [26,28], the size of encoder output is set to 256 or greater. The following network must be wider and deeper to cope with the inflated inputs. As a result, it takes hours to train millions of network parameters, which is not acceptable for fast CT reconstruction.

Frequency-domain encoding is a dense encoder because it utilizes the entire frequency spectrum. However, dense encoding is redundant for CBCT reconstruction for two main reasons. First, a human body usually consists of several homogeneous media, such as muscles and bones. Attenuation coefficients remain approximately uniform inside one medium but vary between different media. High-frequency features are not necessary unless for points near edges. Second, natural objects favor smoothness. Many organs have simple shapes, such as spindle (muscle) or cylinder (bone). Their smooth surfaces can be easily learned with low-dimensional features.

To exploit the aforementioned characteristics of the scanned objects, we use the *hash encoder* [14], a learning-based sparse encoding solution. The equation of hash encoder $\mathcal{M}_{\mathcal{H}}$ is:

$$\mathcal{M}_{\mathcal{H}}(\mathbf{p};\boldsymbol{\Theta}) = [\mathcal{I}(\mathbf{H}_1), \cdots, \mathcal{I}(\mathbf{H}_L)]^T,\ \mathbf{H} = \{\mathbf{c} | h(\mathbf{c}) = (\bigoplus c_j \pi_j) \bmod T\}. \quad (1)$$

Hash encoder describes a bounded space by L multiresolution voxel grids. A trainable feature lookup table $\boldsymbol{\Theta}$ with size T is assigned to each voxel grid. At

each resolution level, we 1) detect neighbouring corners $\mathbf{c}$ (cubes with different colors in Fig. 1(b)) of the queried point $\mathbf{p}$, 2) look up their corresponding features $\mathbf{H}$ in a hash function fashion h [23], and 3) generate a feature vector with linear interpolation $\mathcal{I}$. The output of a hash encoder is the concatenation of feature vectors at all resolution levels. More details of hash function and its symbols can be found in [14].

Compared with frequency encoder, hash encoder produces much smaller outputs (32 in our setting) with competitive feature quality for two reasons. On the one hand, the many-to-one property of hash function conforms to the sparsity nature of human organs. On the other hand, a trainable encoder can learn to focus on relevant details and select suitable frequency spectrum [14]. Thanks to hash encoder, the subsequent network is more compact.

Attenuation Coefficient Prediction. We represent the bounded field with a simple MLP $\mathbf{\Phi}$, which takes the encoded spatial coordinates as inputs and outputs the attenuation coefficients μ at that position. As illustrated in Fig. 1(c), the network is composed of 4 fully-connected layers. The first three layers are 32-channel wide and have ReLU activation functions in between, while the last layer has one neuron followed by a sigmoid activation. A skip connection is included to concatenate the network input to the second layer's activation. By contrast, Zang *et al.* [28] use a 6-layer 256-channel MLP to learn features from a frequency encoder. Our network is 10× smaller.

Attenuation Synthesis. According to Beer's Law, the intensity of an X-ray traversing matter is reduced by the exponential integration of attenuation coefficients on its path. We numerically synthesize the attenuation process with:

$$I = I_0 \exp(-\sum_{i=1}^{N} \mu_i \delta_i), \tag{2}$$

where I_0 is the initial intensity and $\delta_i = \|\mathbf{p}_{i+1} - \mathbf{p}_i\|$ is the distance between adjacent points.

2.3 Model Optimization and Output

NAF is updated by minimizing the L2 loss between real and synthesized projections. The loss function $\mathcal{L}$ is defined as:

$$\mathcal{L}(\mathbf{\Theta}, \mathbf{\Phi}) = \sum_{\mathbf{r} \in \mathbf{B}} \|I_r(\mathbf{r}) - I_s(\mathbf{r})\|^2, \tag{3}$$

where $\mathbf{B}$ is a ray batch, and I_r and I_s are real and synthesized projections for ray $\mathbf{r}$ respectively. We update both hash encoder $\mathbf{\Theta}$ and attenuation coefficient network $\mathbf{\Phi}$ during the training process.

The final output is formulated as a discrete 3D matrix. We build a voxel grid with the desired size and pass the voxel coordinates to the trained MLP to predict the corresponding attenuation coefficients. A CT model thus is restored.

3 Experiments

3.1 Experimental Settings

Data. We conduct experiments on five datasets containing human organ and phantom data. Details are listed in Table 1.

Human Organ: We evaluate our method using public datasets of human organ CTs [4,12], including chest, jaw, foot and abdomen. The chest data are from LIDC-IDRI dataset [4], and the rest are from Open Scientific Visualization Datasets [12]. Since these datasets only provide volumetric CT scans, we generate projections by a tomographic toolbox TIGRE [5]. In TIGRE [5], we capture 50 projections with 3% noise in the range of 180°. We train our model with these projections and evaluate its performance with the raw volumetric CT data.

Phantom: We collect a phantom dataset by scanning a silicon aortic phantom with GE C-arm Medical System. This system captures 582 500 × 500 fluoroscopy projections with position primary angle from −103° to 93° and position secondary angle of 0°. A 512 × 512 × 510 CT image is also generated with inbuilt algorithms as the ground truth. We only use 50 projections for experiments.

Baselines. We compare our approach with four baseline techniques. **FDK** [7] is firstly chosen as a representative of analytical methods. The second method **SART** [2] is a robust iterative reconstruction algorithm. **ASD-POCS** [20] is another iterative method with a total-variation regularizer. We implement a CBCT variant of IntraTomo [28], named **IntraTomo3D**, as an example of frequency-encoding deep learning methods.

Implementation Details. Our proposed method is implemented in PyTorch [17]. We use Adam optimizer with a learning rate that starts at 1×10^{-3} and steps down to 1×10^{-4}. The batch size is 2048 rays at each iteration. The sampling quantity of each ray depends on the size of CT data. For example, we sample 192 points along each ray for the 128 × 128 × 128 chest CT. We use the same hyper-parameter setting for hash encoder as [14]. More details of hyper-parameters can be found in the supplementary material. All experiments are conducted on a single RTX 3090 GPU. We evaluate five methods quantitatively in terms of peak signal-to-noise ratio (PNSR) and structural similarity (SSIM) [25]. PSNR (dB) statistically assesses the artifact suppression performance, while SSIM measures the perceptual difference between two signals. Higher PNSR/SSIM values represent the accurate reconstruction and vice versa.

Table 1. Details of CT datasets used in the experiments.

Dataset name	CT dimension	Scanning method	Scanning range	Number of projections	Detector resolution
Chest [4]	128 × 128 × 128	TIGRE [5]	0°–180°	50	256 × 256
Jaw [12]	256 × 256 × 256	TIGRE [5]	0°–180°	50	512 × 512
Foot [12]	256 × 256 × 256	TIGRE [5]	0°–180°	50	512 × 512
Abdomen [12]	512 × 512 × 463	TIGRE [5]	0°–180°	50	1024 × 1024
Aorta	512 × 512 × 510	GE C-arm	−103°–93°	50 (582)	500 × 500

3.2 Results

Performance. Our method produces quantitatively best results in both human organ and phantom datasets as listed in Table 2. Both PSNR and SSIM values are significantly higher than other methods. For example, the PSNR value of our method in the abdomen dataset is 3.07 dB higher than that of the second-best method **SART**.

We also provide visualization results of different methods in Fig. 2. **FDK** restores low-quality models with notable artifacts, as analytical methods demand large amounts of projections.

Table 2. PSNR/SSIM measurements of five methods on five datasets.

	Chest	Jaw	Foot	Abdomen	Aorta
FDK [7]	22.89/.78	28.59/.78	23.92/.58	22.39/.59	12.11/.21
SART [2]	32.12/.95	32.67/.93	30.13/.93	31.38/.92	27.31/.77
ASD-POCS [20]	29.78/.92	32.78/.93	28.67/.89	30.34/.91	27.30/.76
IntraTomo3D [28]	31.94/.95	31.95/.91	31.43/.91	30.43/.90	29.38/.82
NAF (Ours)	**33.05/.96**	**34.14/.94**	**31.63/.94**	**34.45/.95**	**30.34/.88**

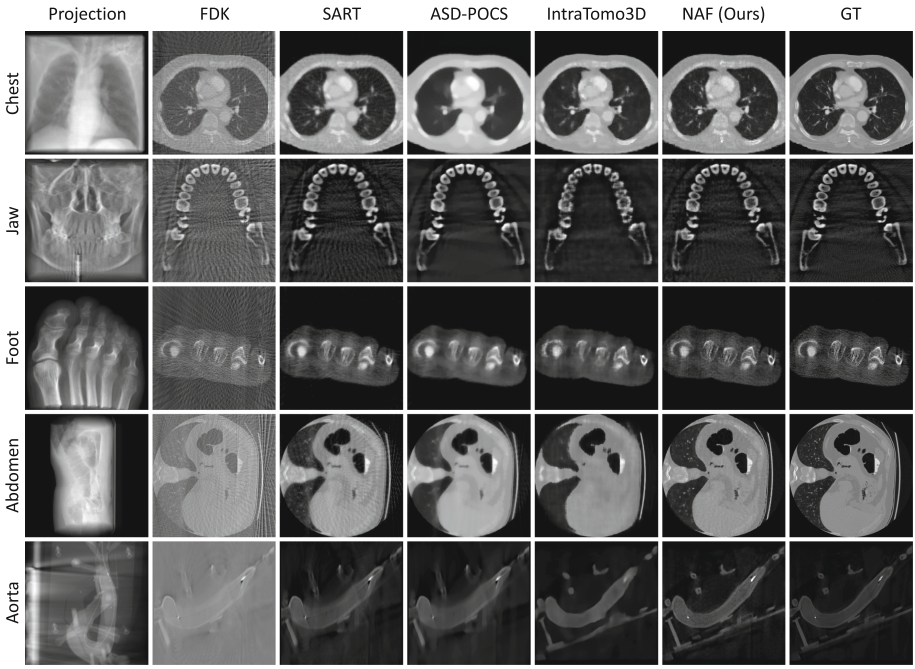

Fig. 2. Qualitative results of five methods. From left to right: examples of X-ray projections, slices of 3D CT models reconstructed by five methods, and the ground truth CT slices.

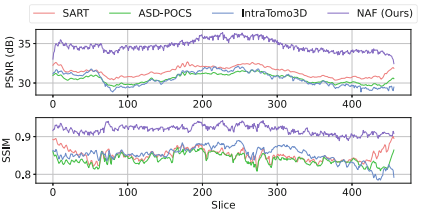

Fig. 3. Slice-wise performance of iterative and learning-based methods on the abdomen dataset.

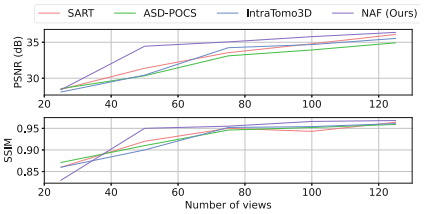

Fig. 4. Performance under different number of views on the abdomen dataset.

Iterative method **SART** suppresses noise in the sacrifice of losing certain details. The reconstruction results of **ASD-POCS** are heavily smeared because total-variation regularization encourages removing high-frequency details, including unwanted noise and expected tiny structures. **IntraTomo3D** produces clean results. However, edges between media are slightly blurred, which shows that the frequency encoder fails to teach the network to focus on edges. With the help of hash encoding, results of the proposed **NAF** have the most details, clearest edges and fewest artifacts. Figure 3 indicates that **NAF** outperforms other methods in all slices of the reconstructed CT volume.

Figure 4 shows the performance of iterative methods and learning-based methods under different number of views. It is clear that the performance increases with the rise of input views. Our methods achieves better results than others under most circumstances.

Time. We record the running time of iterative and learning-based methods as shown in Fig. 5. All methods use CUDA [15] to accelerate the computation process. Overall, the methods spend less time on datasets with small projections (chest, jaw and foot) and increasingly more time on big datasets (abdomen and aorta). **IntraTomo3D** requires more than one hour to train the network. Benefiting from the compact network design, **NAF** spends similar running time to iterative methods and is 3× faster than the frequency-encoding deep learning method **IntraTomo3D**.

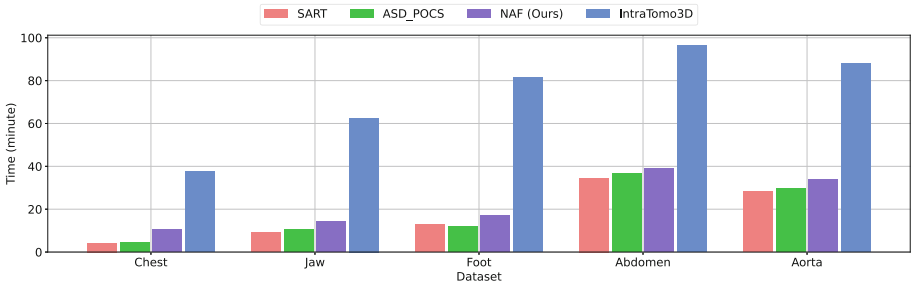

Fig. 5. Running time that iterative and learn-based methods take to converge to stable results.

4 Conclusion

This paper proposes NAF, a fast self-supervised learning-based solution for sparse-view CBCT reconstruction. Our method trains a fully-connected deep neural network that consumes a 3D spatial coordinate and outputs the attenuation coefficient at that location. NAF synthesizes projections by attenuating incident X-rays based on the predicted attenuation coefficients. The network is updated by minimizing the projection error. We show that frequency encoding is not computationally efficient for tomographic reconstruction tasks. As an alternative, a learning-based encoder entitled hash encoding is adopted to extract valuable features. Experimental results on human organ and phantom datasets indicate that the proposed method achieves significantly better results than other baselines and spends reasonably short computation time.

References

1. Adler, J., Öktem, O.: Learned primal-dual reconstruction. IEEE Trans. Med. Imaging **37**(6), 1322–1332 (2018)
2. Andersen, A.H., Kak, A.C.: Simultaneous algebraic reconstruction technique (SART): a superior implementation of the art algorithm. Ultrason. Imaging **6**(1), 81–94 (1984)
3. Anirudh, R., Kim, H., Thiagarajan, J.J., Mohan, K.A., Champley, K., Bremer, T.: Lose the views: limited angle CT reconstruction via implicit sinogram completion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6343–6352 (2018)
4. Armato III, S.G., et al.: The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans. Med. Phys. **38**(2), 915–931 (2011)
5. Biguri, A., Dosanjh, M., Hancock, S., Soleimani, M.: TIGRE: a MATLAB-GPU toolbox for CBCT image reconstruction. Biomed. Phys. Eng. Express **2**(5), 055010 (2016)
6. Chen, H., et al.: Learned experts' assessment-based reconstruction network ("learn") for sparse-data CT, arXiv preprint arXiv:1707.09636 (2017)
7. Feldkamp, L.A., Davis, L.C., Kress, J.W.: Practical cone-beam algorithm. Josa a **1**(6), 612–619 (1984)
8. Gao, Y., et al.: Low-dose x-ray computed tomography image reconstruction with a combined low-mas and sparse-view protocol. Opt. Express **22**(12), 15190–15210 (2014)
9. Hornik, K., Stinchcombe, M., White, H.: Multilayer feedforward networks are universal approximators. Neural Netw. **2**(5), 359–366 (1989)
10. Kang, E., Chang, W., Yoo, J., Ye, J.C.: Deep convolutional framelet denosing for low-dose CT via wavelet residual network. IEEE Trans. Med. Imaging **37**(6), 1358–1369 (2018)
11. Kasten, Y., Doktofsky, D., Kovler, I.: End-To-end convolutional neural network for 3D reconstruction of knee bones from Bi-planar X-ray images. In: Deeba, F., Johnson, P., Würfl, T., Ye, J.C. (eds.) MLMIR 2020. LNCS, vol. 12450, pp. 123–133. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-61598-7_12

12. Klacansky, P.: Open scientific visualization datasets (2022). http://klacansky.com/open-scivis-datasets/
13. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: NeRF: representing scenes as neural radiance fields for view synthesis. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020, Part I. LNCS, vol. 12346, pp. 405–421. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58452-8_24
14. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. arXiv:2201.05989, January 2022
15. NVIDIA, Vingelmann, P., Fitzek, F.H.: Cuda, release: 10.2.89 (2020). http://developer.nvidia.com/cuda-toolkit
16. Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: DeepSDF: learning continuous signed distance functions for shape representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 165–174 (2019)
17. Paszke, A., et al.: Pytorch: An imperative style, high-performance deep learning library. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 32, pp. 8024–8035. Curran Associates, Inc. (2019). http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf
18. Rahaman, N., et al.: On the spectral bias of neural networks. In: International Conference on Machine Learning, pp. 5301–5310. PMLR (2019)
19. Scarfe, W.C., Farman, A.G., Sukovic, P., et al.: Clinical applications of cone-beam computed tomography in dental practice. J. Can. Dent. Assoc. **72**(1), 75 (2006)
20. Sidky, E.Y., Pan, X.: Image reconstruction in circular cone-beam computed tomography by constrained, total-variation minimization. Phys. Med. Biol. **53**(17), 4777 (2008)
21. Tancik, M., et al.: Fourier features let networks learn high frequency functions in low dimensional domains. Adv. Neural Inf. Process. Syst. **33**, 7537–7547 (2020)
22. Tang, C., et al.: Projection super-resolution based on convolutional neural network for computed tomography. In: 15th International Meeting on Fully Three-Dimensional Image Reconstruction in Radiology and Nuclear Medicine, vol. 11072, p. 1107233. International Society for Optics and Photonics (2019)
23. Teschner, M., Heidelberger, B., Müller, M., Pomerantes, D., Gross, M.H.: Optimized spatial hashing for collision detection of deformable objects. In: VMV, vol. 3, pp. 47–54 (2003)
24. Wang, C., et al.: Improving generalizability in limited-angle Ct reconstruction with sinogram extrapolation. In: de Bruijne, M., et al. (eds.) MICCAI 2021, Part VI. LNCS, vol. 12906, pp. 86–96. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87231-1_9
25. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
26. Wu, Q., et al.: IREM: high-resolution magnetic resonance image reconstruction via implicit neural representation. In: de Bruijne, M., et al. (eds.) MICCAI 2021, Part VI. LNCS, vol. 12906, pp. 65–74. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87231-1_7

27. 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)
28. Zang, G., Idoughi, R., Li, R., Wonka, P., Heidrich, W.: IntraTomo: self-supervised learning-based tomography via sinogram synthesis and prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1960–1970 (2021)