IOP Publishing Phys. Med. Biol. 69 (2024) 045016 https://doi.org/10.1088/1361-6560/ad1d6c

Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

RECEIVED
28 September 2023

REVISED
27 December 2023

ACCEPTED FOR PUBLICATION
11 January 2024

PUBLISHED
8 February 2024

PAPER

ACnerf: enhancement of neural radiance field by alignment and correction of pose to reconstruct new views from a single x-ray*

Mengcheng Sun[1], **Yu Zhu**[1,**], **Hangyu Li**[1], **Jiongyao Ye**[1] and **Nan Li**[2,**]

[1] School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China, People's Republic of China

[2] Dapartment of Orthopedics, 96603 Military Hospital of PLA, Huaihua 418000, People's Republic of China

** Authors to whom any correspondence should be addressed.

E-mail: zhuyu@ecust.edu.cn **and** 958695747@qq.com

Keywords: neural radiance field, x-ray, computed tomography, rendering, projection

Abstract

Objective. Computed tomography (CT) is widely used in medical research and clinical diagnosis. However, acquiring CT data requires patients to be exposed to considerable ionizing radiance, leading to physical harm. Recent studies have considered using neural radiance field (NERF) techniques to infer the full-view CT projections from single-view x-ray projection, thus aiding physician judgment and reducing Radiance hazards. This paper enhances this technique in two directions: (1) accurate generalization capabilities for control models. (2) Consider different ranges of viewpoints. *Approach.* Building upon generative radiance fields (GRAF), we propose a method called ACnerf to enhance the generalization of the NERF through alignment and pose correction. ACnerf aligns with a reference single x-ray by utilizing a combination of positional encoding with Gaussian random noise (latent code) obtained from GRAF training. This approach avoids compromising the 3D structure caused by altering the generator. During inference, a pose judgment network is employed to correct the pose and optimize the rendered viewpoint. Additionally, when generating a narrow range of views, ACnerf employs frequency-domain regularization to fine-tune the generator and achieve precise projections. *Main results.* The proposed ACnerf method surpasses the state-of-the-art NERF technique in terms of rendering quality for knee and chest data with varying contrasts. It achieved an average improvement of 2.496 dB in PSNR and 41% in LPIPS for 0°–360° projections. Additionally, for −15° to 15° projections, ACnerf achieved an average improvement of 0.691 dB in PSNR and 25.8% in LPIPS. *Significance.* With adjustments in alignment, inference, and rendering range, our experiments and evaluations on knee and chest data of different contrasts show that ACnerf effectively reduces artifacts and aberrations in the new view. ACnerf's ability to recover more accurate 3D structures from single x-rays has excellent potential for reducing damage from ionising radiation in clinical diagnostics.

1. Introduction

The 3D medical data generated by technologies such as computed tomography (CT) and magnetic resonance imaging (MRI) often provide adequate visual information to assist physicians in diagnosis (Suetens 2009). However, this is often accompanied by high costs. In the case of CT data, for example, the basic principle is to scan a certain thickness of layers of the body's examination area with an x-ray beam to produce multiple slices and overlap the slice information to obtain 3D data, but this requires prolonged exposure of the patient to a higher level of Radiance compared to a single x-ray image (Lo *et al* 2012). Physicians usually choose to judge 3D information from a few x-ray images for cost-effectiveness, but this relies heavily on human *a priori* knowledge.

* This work was supported in part by the Science and Technology Commission of Shanghai Municipality under Grant 20DZ2254400 and 20DZ2261200; in part by Shanghai Municipal Science and Technology Major Project under Grant ZD2021CY001.

A practical challenge is reconstructing projections of 3D CT data from a small number of 2D x-rays using computer techniques (Kasten *et al* 2020).

In the context of a given imaging system, early approaches involved establishing mathematically compatible models that iteratively and analytically reconstructed the 3D information of medical images (Huynh *et al* 2015, Xie *et al* 2020). However, its application is greatly restricted when the imaging system is unknown or incompatible with the mathematical model. With the development of deep learning, much research has been focused on utilizing sparse views to reconstruct three-dimensional CT data from a limited number of two-dimensional medical images (Li *et al* 2019, Lindell *et al* 2021, Sun *et al* 2021, Shen *et al* 2022, Cheng *et al* 2023). These methods overcome the unknowns in the imaging system and the mismatch of mathematical models, but they require annotated data and supervised paired 3D data. Consequently, the expensive annotation and training costs make it challenging to generalize these methods to niche medical fields.

Recently, neural radiance field (NERF) has received much attention in medicine, which can be used to estimate an implicit representation of its 3D structure using 2D images by neural networks (Mildenhall *et al* 2021). The neural network can query the density and color information of the points on the novel view ray, and then using Volume rendering, the new view can be drawn. This technique is proposed with stringent requirements. Firstly, the training image scene must be stationary, or the images can be assumed to have been taken simultaneously. Secondly, the number of training images largely determines the effectiveness of the novel view synthesis (Martin-Brualla *et al* 2021, Yu *et al* 2021b). Also, capturing multiple patient images in a short time is difficult and inconsistent with economic and health assumptions.

To overcome the shortcomings of the original NERF, some models adopt a sparse view and leverage prior knowledge to generate feature maps (Wang *et al* 2021, Yu *et al* 2021b), geometric depth information (Chen *et al* 2021), and adversarial training to endow the radiance field with generalization capability (Schwarz *et al* 2020, Trevithick and Yang 2021). Although these models have achieved competitive results on natural images, they have not made targeted improvements for medical images, particularly in cases where the imaging scene and initial angles are fixed. One of them, GRAF (Schwarz *et al* 2020), uses inputs from random views and an adversarial supervised training scheme, which enables it to excel in generalization ability and image re-editing. MEDnerf (Corona-Figueroa *et al* 2022) has improved GRAF by attempting to restore the complete CT projection with single-view x-ray, which revealed the potential of the NERF technique in this direction. Fine-tuning the GRAF model's generator alone for creating complete CT projections of specific patients does not meet the accuracy requirements of medical imaging due to limited generalization capabilities.

While investigating the ability of NERF models to generalize over medical images, we observed an intriguing phenomenon. The process of fine-tuning the model generator using solely a single reference x-ray of a specific patient resulted in artifacts, along with significant distortions, in the other generated 3D CT projections, as shown in figure 1. MEDnerf. These issues stem from the destruction of the trained implied 3D structure. In this paper, to mitigate this, we propose the ACnerf, which combines latent code with position encoding to propose a linear form of coding that fully utilizes its editing potential during the fine-tuning phase. At the same time, we design a pose judgment network to correct the error between the output image and its corresponding pose, as shown in figure 2. Furthermore, inspired by FREEnerf (Yang *et al* 2023), we design a frequency-domain regularization to fine-tune the generator to reconstruct small-range projections. In contrast to state-of-the-art NERF methods, comprehensive experiments on full 360° and small-range predictions on keen (Ali *et al* 2016) and chest (Clark *et al* 2013) datasets show that ACnerf offers significant advantages across a wide range of challenges. The main contributions are summarized below:

(1) We propose a solution to enhance the quality of generated images through alignment. Our approach utilizes latent encoding to linearly impact positional encoding, enabling the generator to accurately perform alignment operations without disturbing the data manifold captured during the training phase.

(2) We designed a pose judgment network to perform pose correction operations. The network introduces additional prior knowledge of medical images into the NERF, which effectively reduces the pose error.

(3) The proposed ACnerf achieves better results on the experimental datasets. Compared to the state-of-the-art model, the Peak Signal-to-Noise Ratio (PSNR) is improved 2.452 dB, 2.54 dB and the Learned Perceptual Image Patch Similarity (LPIPS) is improved by 44%, 38% on knee and chest for the 360° views, respectively.

(4) Frequency-domain regularization scheme is employed to optimize the generator, enhancing projection accuracy for small-range reconstruction and rendering. In the −15° to 15° views, the PSNR is improved by 0.767 dB, 0.615 dB and the LPIPS is improved by 34% and 11%, respectively.

2. Related work

In recent years, NERF have achieved significant breakthroughs across various areas, garnering growing attention. This chapter will review the NERF technology and discuss critical research on sparse viewpoint and

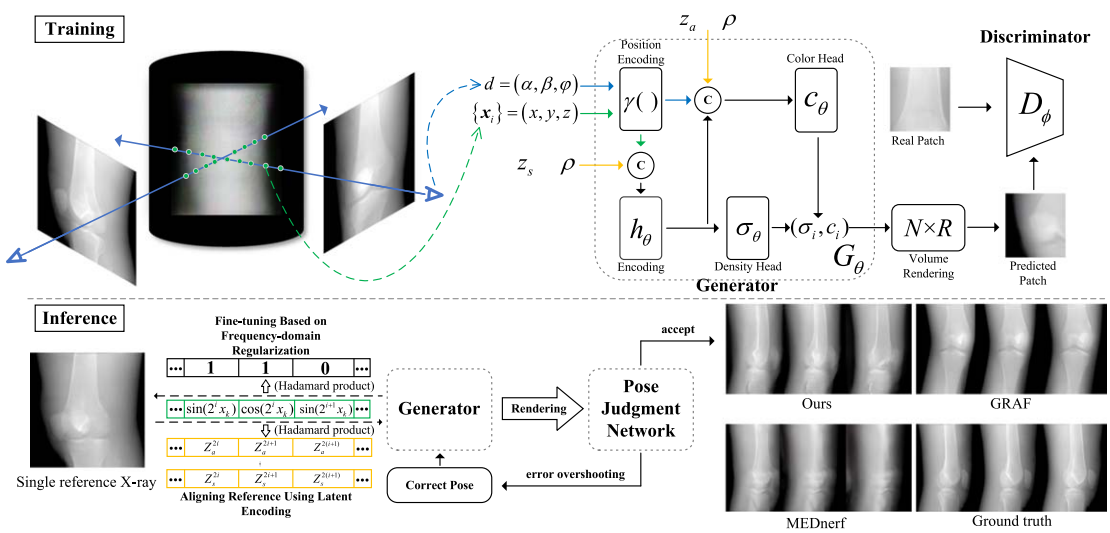

Figure 1. The total process for reasoning about new projections from a single x-ray using NERF technology. The first half is the training phase of the model, where the generator infers the density and color of the particles based on their position in space and the observation pose. The *N* particles in a ray are volume rendering to get a pixel value, and the pixel values obtained from the *R* rays form a patch. Position information is conveyed using green arrows, pose information is in blue, and latent code introduced to control the network is in yellow. The second half is the inference phase, which involves aligning the network with the reference pose by fine-tuning it, and then changing the input pose of the network to get a new projection.

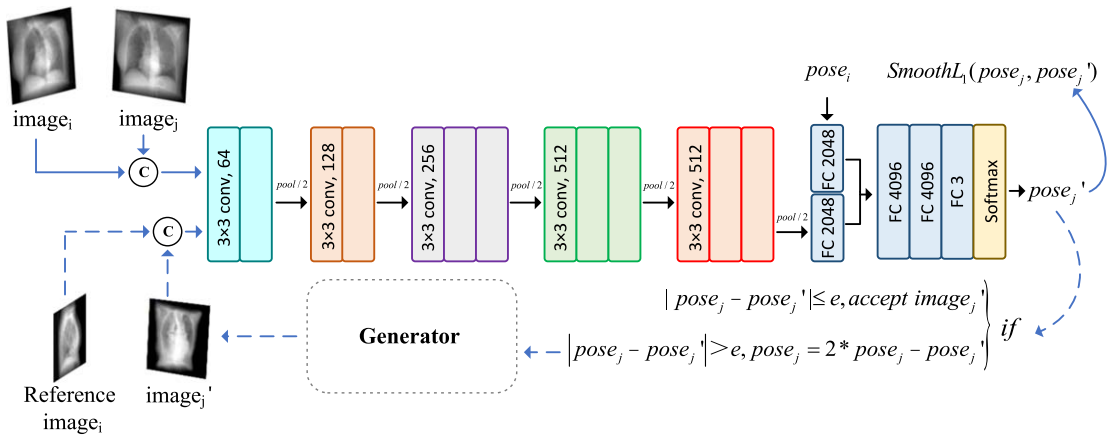

Figure 2. Flowchart of the pose judgment network. The blue solid and dashed parts are the inputs and outputs of the training and inference phases, respectively, and the pose is unified with 3D Euler angles. When the error between the network's predicted *pose′* and the generator's input *pose* is less than *e*, the predicted projected *image′* is accepted. Otherwise, the *pose* is updated to continue the generator's inference.

generalization capabilities. Finally, we will report on the application of NERF technology in medical image reconstruction.

2.1. Neural radiance field

In the representation of 3D structure, implicit representations are more adept at handling topological structures and describing relationships between points (Xie *et al* 2022). As powerful function approximators, neural networks hold an advantage in implicit representation schemes for scenes. The idea behind NERF is to represent the target reconstructed 3D structure as a continuous function parameterized by neural networks (Mildenhall *et al* 2021). The foundational research of NERF primarily focuses on enhancing the reconstruction quality through rendering viewpoints (Barron *et al* 2021, Martin-Brualla *et al* 2021, Barron *et al* 2022, Roessle *et al* 2023), accelerating training and inference (Liu *et al* 2020, Park *et al* 2021, Yu *et al* 2021a, Müller *et al* 2022, Chen *et al* 2023a), and addressing dynamic scene deformations (Park *et al* 2021, Xu and Harada 2022, Liu *et al* 2023b). Currently, NERF has found wide applications in various domains, including 3D editing tasks (Jain *et al* 2022, Poole *et al* 2022), segmentation tasks (Ranade *et al* 2022, Cen *et al* 2023, Siddiqui *et al* 2023), and facial

reconstruction (Guo *et al* 2021, Isık *et al* 2023, Wang *et al* 2023), among others. One strict condition in these tasks is the need for a large number of images to be used as a reference to obtain a high-quality representation of the scene, which significantly limits its use in real-life applications.

2.2. Few-shot and zero-shot NERF

Many studies attempt to introduce additional information to address the problem of few-shot volume rendering. Leveraging pre-trained networks' prior knowledge to generate feature maps (Wang *et al* 2021, Yu *et al* 2021b), incorporating depth information (Chen *et al* 2021), or incorporating specific scene geometrical priors (Kulhánek *et al* 2022, Niemeyer *et al* 2022, Yang *et al* 2023) can all alleviate the issue to some extent. With the improvement of computational power, some studies have incorporated diffusion models from image generation tasks into the training of NERF, showing excellent performance in completely unknown single-view reconstruction tasks (Liu *et al* 2023a, Wynn and Turmukhambetov 2023). Although pre-trained diffusion models can generate various images, precise control over the viewing pose becomes challenging, while expensive inference and training costs are also required. In constrained categories, NERF with generalization capability has an advantage. The network can capture the manifold of specific types rather than a single scene by employing adversarial training (Schwarz *et al* 2020, Trevithick and Yang 2021), random input position encoding, and pose. However, these methods often struggle with registration in medical scenarios that demand high precision and accuracy regarding angles.

2.3. NERF in medicine

Currently, the medical field is primarily focused on exploring the potential of NERF technology in 3D reconstruction and has achieved some initial progress. When multiple reference projections (more than 50) are available, NAF (Zha *et al* 2022) restores 3D CT data by modifying the rendering technique of radiation fields. SNAF (Fang *et al* 2022) further enhances data quality and reduces the input of projections (requiring over 30) by utilizing a pre-trained denoising module. DIF-net (Lin *et al* 2023) adopts U-net as the feature extraction network and uses the feature information from the reference projections as input for the radiation field. It utilizes 3D data for supervised training, but still necessitates the reconstruction of five or more projections. When 3D data is available, Cunerf (Chen *et al* 2023b) utilizes a voxel-based sampling and rendering method to achieve zero-shot super-resolution reconstruction. Ultra-nerf (Wysocki *et al* 2023) applies radiation field techniques to medical scenarios with continuous multi-view references, enabling new view reconstruction of ultrasound videos. MEDnerf (Corona-Figueroa *et al* 2022) considers a more realistic scenario where a single x-ray image is obtained as a reference and uses a radiation field model with generalization capability to reconstruct CT projections from new views. Still, MEDnerf can be difficult to align and produce distortions when generating continuous CT projections of the patient, which remains a challenging problem.

3. Method

In this section, we first present the overall flow of inferring new views from a single x-ray image, as shown in figure 1. In the overall flow, we describe the training process using different views of x-ray images with GRAF (Schwarz *et al* 2020) as the backbone, and briefly describe how the inference process can be optimized by two optimization schemes, alignment and pose correction. Subsequently, we describe in detail the workflow and application scope of these two schemes. Finally, we propose a fine-tuning scheme to accurately generate projections on a small range. The specific implementation details of these contents are as follows:

3.1. Network overview

As shown in figure 1 (Training), The NERF establishes a continuous mapping from the position of points $\boldsymbol{x} = (x, y, z)$ and the viewing pose d to the color content c and volume density σ of the points, thus representing the implicit 3D structure. Specifically, when the color content c and volume density σ of the particles traversed by the ray $r(t) = o + td$ emitted along the direction d from the center of projection o in space are known, the pixel color $C(r)$ can be computed using volume rendering:

$$C(r) = \int_{t_n}^{t_f} \frac{\sigma(r(t))c(r(t),\, d)dt}{exp(\int_{t_n}^{t} o(r(s))ds)}, \tag{1}$$

where $c(\bullet)$ and $\sigma(\bullet)$ represent the computation functions for color and volume density, given the sampled results of N points (from t_n to t_f), an approximate volume rendering integral $\hat{C}(r)$ can be computed as follows (Max 1995):

$$\hat{C}(r) = \sum_{i=1}^{N} \frac{1 - \exp(-\sigma_i(t_{i+1} - t_i))}{\exp(\sum_{j=1}^{i} \sigma_j(t_{j+1} - t_j))} c_i. \tag{2}$$

In practical applications, MLP networks are commonly used to approximate $c(\bullet)$ and $\sigma(\bullet)$:

$$(c_i, \sigma_i) = F_{MLP}(\gamma(\boldsymbol{x_i}), \gamma(d)), \tag{3}$$

where $\gamma(\boldsymbol{x_i})$ and $\gamma(d)$ represent the high-frequency encodings of point positions and observation directions, the process of high-frequency encoding can be described as follows:

$$\gamma(p) = p \bigcup_{i=0}^{L-1} (\sin(2^i p), \cos(2^i p)), \quad L \in N. \tag{4}$$

Using patches of $K \times K$ rays selected with random poses as inputs for the discriminator D_ϕ. It adopts the unsaturated GAN loss (Mescheder *et al* 2018) to bring the distribution of generated images closer to ground truth:

$$\begin{aligned} L(\theta, \phi) = E_{z_s, z_a \sim N(\vec{0}, I)}[f(D_\phi(G_\theta(Z_s, Z_a, \xi, \nu)))] \\ + E(f(-D_\phi(I) + \lambda ||\nabla D_\phi(I)||^2)), f(t) = -\log(1 + \exp(-t)), \end{aligned} \tag{5}$$

Where ξ and ν are sets of points and poses that render the patch, respectively. Z_a and Z_s are 128-dimensional latent codes randomly sampled from a Gaussian distribution during training, used to fine-tune the appearance and shape of specific images. After training, a new projection is obtained by changing the inputs after aligning the model with the reference image as shown in figure 1 (Inference). The adversarial training approach allows GRAF to be unfamiliar with the ground truth pose, leading to errors in the new projections (figure 1 GRAF). To alleviate this problem, MEDnerf employs a relaxed reconstruction formulation (Pan *et al* 2021) to align the reference projection:

$$\theta^*, Z_s^*, Z_a^* = \underset{\theta, Z_a, Z_s}{\operatorname{argmin}} L(I, G(Z; \theta)), \quad I^* = G(Z^*; \theta^*), \tag{6}$$

where L denotes the mean square error, and the optimal θ^* and Z^* are sought by fine-tuning θ and Z so that the projection I^* rendered by the generator is close to the reference projection I at the same pose. In single-view reconstruction, the generator G is highly susceptible to overfitting, resulting in only high-quality images of the reference angle and distortions at unknown angles (as shown in figure 1 MEDnerf), which we do not expect. Therefore, during the inference stage (as shown in figure 1 Inference), we devised an approach to integrate the latent encoding Z with positional encoding. This approach involves adjusting Z to influence the sampling positions, thereby aligning the network with the reference x-ray while avoid modifying the generator's parameters G_θ. Following the alignment, to mitigate errors in new views caused by pose inaccuracies, we designed the pose judgment network. It evaluates the pose of the rendered new view and accepts it when the pose meets the predefined criteria (as shown in figure 1 (Ours)). Otherwise, the input to the generator is modified. We provide detailed explanations of the aforementioned reference alignment and pose correction methods in sections 3.2 and 3.3, respectively.

3.2. Aligning reference projection using latent encoding

In GRAF, the observation angle, particle position, appearance, and shape of the object are disentangled, with Z_s and Z_a separately controlling shape and appearance. The capability is limited when aligning the network to a specific space solely by adjusting Z_s and Z_a. In scenarios where image editing is not required, a certain level of coupling in the system is permissible. With this consideration, we design latent encoding to linearly influence position encoding, following the process outlined below:

$$\gamma(\boldsymbol{x})^* = Z_s \odot \gamma(\boldsymbol{x}) + Z_a, \tag{7}$$

where $\gamma(\boldsymbol{x})^*$ represents the adjusted position encoding, $\odot$ signifies the Hadamard product operation, and $+$ denotes the concat operation. By employing this method to extend the influence range of $\gamma(\boldsymbol{x})$, the reconstruction formula transforms into:

$$Z_s^*, Z_a^* = \underset{Z_s, Z_a}{\arg\min} L(I, G(Z; \gamma(\boldsymbol{x}))), \quad I^* = G(Z^*; \gamma(\boldsymbol{x})^*). \tag{8}$$

Throughout this process, the generator's parameters have not been altered, thereby avoiding the image distortions caused by equation (6). In the experimental section 4.3.1, we demonstrate that the convergence direction of equation (8) aligns with the enhancement direction of projection quality for unknown poses.

3.3. Pose judgment network

In order to achieve generalization across different image spaces, the GRAF does not include rigorous training for the pose. Consequently, precise correction is required, despite the model's continuous understanding of the pose, particularly when generating images specific to individual patients. Building upon aligning the input poses to the reference image in section 3.2, we design a pose judgment network to correct errors in the projections generated under unknown poses, as shown in figure 2. The backbone of the network employed in the study utilizes VGG16. The inputs consist of two CT projections acquired from the same patient but with different poses, along with the pose information corresponding to one of the projections. The primary objective of the network is to predict the pose of the other projection. In the inference stage, the reference image, the generating image, and the reference pose are used as inputs to get the pose judged by the network, and the specific process can be expressed as follows:

$$\alpha_j', \beta_j', \varphi_j' = F_{vgg16}(I_i, I_j, \alpha_i, \beta_i, \varphi_i), \tag{9}$$

where α, β, φ represents the three-dimensional pose of the projection. When the predicted pose is close to the input of the generator, the output of the generator is accepted. Otherwise, the input to the generator is adjusted, following the strategy outlined below:

$$\begin{aligned} v - (F_{vgg16}(I_R, I_G)) &\leqslant e \\ I_G = G_\theta(Z_s, Z_a, \xi, 2v - F_{vgg16}(I_R, I_G)), & \end{aligned} \tag{10}$$

Where I_R and I_G represent the reference projection and generated projection, respectively. When the difference between the predicted view angle and the input view angle is less than e, the generated image is accepted. Otherwise, the network input view pose is corrected. During inference, the reference projection's pose is chosen as the initial, and a 360° projection is generated with it as the center.

3.4. Network fine-tuning based on frequency-domain regularization

During the fine-tuning process, altering the generator's parameters carries a significant risk of inducing distortion in the projections, particularly when dealing with large angles. When relying solely on a reference single x-ray, a more practical application is to obtain accurate images for a small-range. In FREEnerf (Yang *et al* 2023), a frequency-domain regularization approach is employed to train under sparse views, aiming to prevent network overfitting caused by high-frequency position encoding. Taking inspiration from this approach, we build upon section 3.2 and apply a frequency-domain regularization to the position encoding in the network's input. We fine-tune the generator's parameters, gradually reducing the high-frequency mask as the training epochs progress. This process can be described as follows:

$$\begin{aligned} \gamma(\boldsymbol{x})' &= \gamma(\boldsymbol{x}) \odot \alpha(t), \\ \text{with } \alpha_i(t, T, L) &= \begin{cases} 1, & \text{if } i \leqslant \frac{t \cdot L}{T} + 39 \\ 0, & \text{otherwise} \end{cases}, \end{aligned} \tag{11}$$

where $\alpha(t)$ represents the position of the mask, T is the total number of training epochs, and as the epochs t increase, the mask is gradually released following the amplitude L. The generator's parameters are fine-tuned through mean squared error between the generated and reference images. When the network focuses more on the low-frequency region, inputs with small angles tend to align, thereby alleviating mode collapse. During inference, we generate projections ranging from −15° to 15°around the pose of the reference to validate their quality.

4. Experiment

4.1. Dataset and metrics

Collecting paired x-ray images and CT data can lead to errors due to patient movement and equipment variations, and it can also subject patients to higher Radiance exposure. Therefore, following the approach of MEDnerf, we employ digital radiographic radiography (DRR) techniques to obtain simulated projections from 5 knee joint CT datasets (Ali *et al* 2016) and 20 chest CT datasets (Clark *et al* 2013). These datasets encompass patients with varying contrasts. During the simulation of projections, we assume that the x-ray source and the projection panel are parallel. Projections are generated at 5-degree intervals around the Z-axis, resulting in 72 angles for each CT dataset. The resolution is set to 128 × 128. During training, we randomly sample 80% of the complete patient data, which includes chest data from 16 patients and knee data from 4 patients. During testing, we provide a random view from the remaining patients as a reference and generate projections for the remaining 71 views. This work does not involve experimental procedures with human subjects or animals.

Table 1. Quantitative comparisons in 0° to 360° were evaluated using PSNR ↑, SSIM ↑, and LPIPS ↓. The best and second-best results are marked in red and blue, respectively. Our direct baseline is GRAF.

Method	knee dataset			chest dataset		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
PIXELnerf	17.059	0.529	0.469	16.874	0.335	0.432
GRAF	14.717	0.524	0.360	15.493	0.328	0.337
MEDnerf	15.356	0.538	0.311	15.524	0.339	0.326
FREEnerf	12.641	0.308	0.434	13.173	0.288	0.621
our w/o PJ	17.194	0.563	0.206	17.728	0.350	0.215
ours	17.808	0.584	0.174	18.064	0.447	0.202

We quantitatively evaluate the results based on three visual metrics: Peak Signal-to-Noise Ratio (PSNR), Structural SimilarityS (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS) (Zhang *et al* 2018). PSNR and SSIM focus more on pixel-level differences, while LPIPS measures perceptual differences that are highly correlated with human perception. The higher the similarity between images, the closer the PSNR value is to infinity, the closer the SSIM value is to 1, and the closer the LPIPS value is to 0. In the experiments, all models are built using the PyTorch framework and run on a single NVIDIA RTX A6000 GPU with 48GB of memory. For a fair comparison, we maintain uniform hyperparameters across all models: position encoding $\gamma(\boldsymbol{x}) = 63$, pose encoding $\gamma(d) = 27$, utilizing the Adam optimizer with a learning rate annealing from 10^{-4} to 10^{-6} through cosine annealing, a batch size of 16, and a total of 100 000 iterations. When aligning the generalization models (Ours, GRAF, MEDnerf) to the images, we uniformly employ PSNR as the output metric, considering the alignment complete when the increase in PSNR is less than 0.5% in every 50 iterations. The pose error limits $e = 0.04$. Due to the lack of generalization in FREEnerf (Yang *et al* 2023), it cannot utilize the training set. We provide 18 views during testing to establish a comparable benchmark with other models.

4.1.1. Frequency-domain regularization

Table 5 presents the rendering results of different methods after fine-tuning the generator using approach 3.4, all of which were already aligned before the fine-tuning process. It can be observed that within a small-range,

4.2. Comparison with state-of-the-art models

We compare the proposed ACnerf with four state-of-the-art methods, including two models with excellent generalization capabilities and two models specifically designed for sparse view reconstruction. Among them, GRAF (Schwarz *et al* 2020) enhances the generalization of the original radiance field by introducing random noise and utilizing adversarial learning mechanisms during the training phase. In the inference phase, random noise is treated as a code and modified to edit the projection. Building upon this, MEDnerf (Corona-Figueroa *et al* 2022) further optimizes the training process by subjecting the patches generated by GRAF to three enhancement schemes with a discriminator, resulting in visually improved images. PIXELnerf (Yu *et al* 2021b) leverages a pre-trained classification network (Resnet34) to extract features from reference images and combines them with the spatial coordinates of the radiance field as inputs to the network. This incorporation of additional prior knowledge helps alleviate the information deficiency caused by sparse views. On the other hand, FREEnerf (Yang *et al* 2023) considers reasonable constraints and employs frequency-domain regularization in the training process to prevent overfitting caused by sparse views, but it lacks generalization capabilities. Based on different practical needs, the experiments investigate the quality of 360° projections and small-range projection quality.

4.2.1. Quantitative comparison

Table 1 reports the performance of various models at 0° to 360°. For generating 360° projections, to prevent distortion in distant projections caused by changes in generator parameters, two approaches proposed in sections 3.2 (ours w/o PJ) and 3.3 (ours) are progressively adopted. We can observe that the projection quality of both schemes of ACnerf outperforms all competitors, especially the LPIPS, which is highly correlated with image perception. The PSNR performance of PIXELnerf (Yu *et al* 2021b) is closer to ours on the knee, but not on the chest, which has a more complex organizational structure. Other schemes that focus on training patterns without introducing *a priori* information during training all perform better on the chest. This confirms the correctness of our motivational direction. Table 2 shows the performance of the different models at −15° to 15°. In the localized projection (−15° to 15°), we employ the fine-tuning strategy from section 3.4 (ours+) based on the foundation of section 3.2. In this aspect, we emphasize pixel-level accuracy more strongly, and ACnerf continues to maintain the highest projection quality. Compared to PIXELnerf (Yu *et al* 2021b) and GRAF

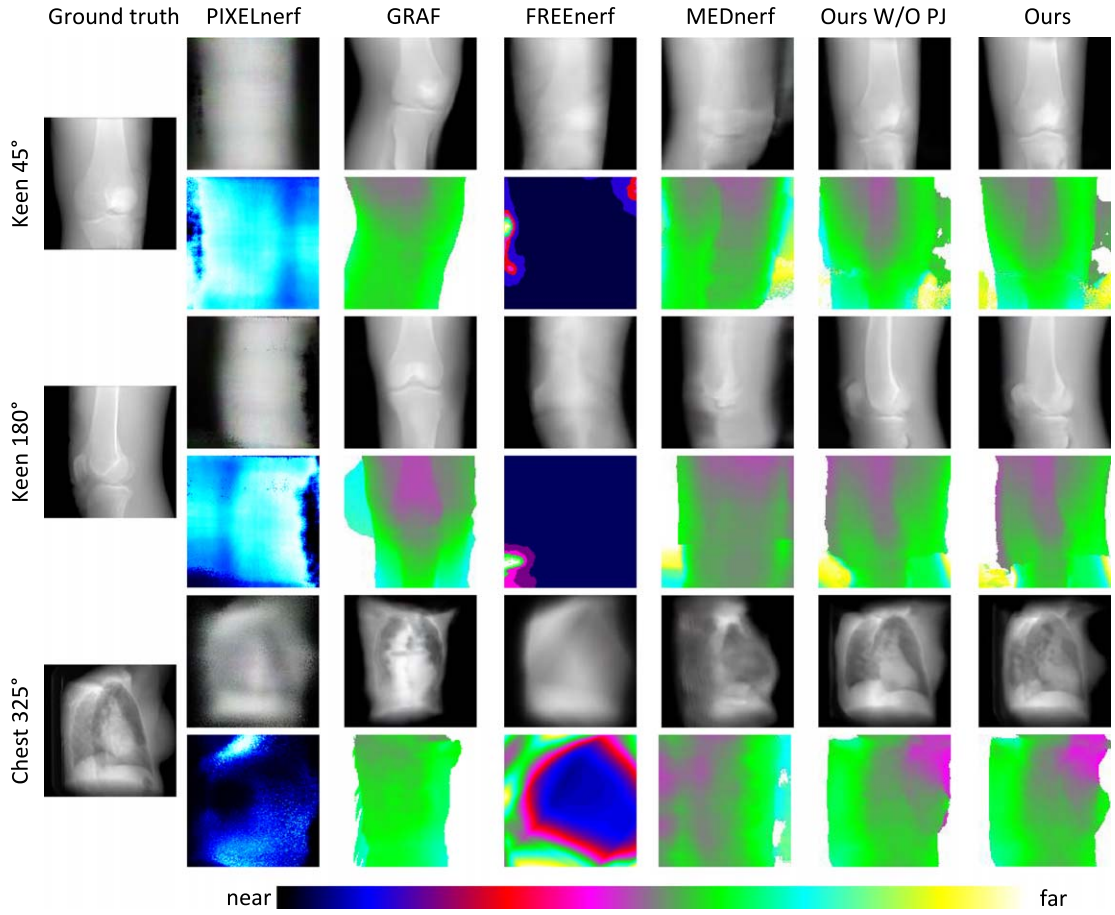

Figure 3. Comparison of visualization of 0°–360° projections. Rows 2, 4, and 6 are depth maps representing the distribution of particles in space.

Table 2. Quantitative comparisons at −15° to 15° and −5° to 5° were evaluated using PSNR ↑, SSIM ↑, and LPIPS ↓. The best and second-best results are marked in red and blue, respectively.

range	Method	knee dataset			chest dataset		
		PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
-15° to 15°	PIXELnerf	19.469	0.569	0.542	18.551	0.488	0.582
	GRAF	18.341	0.625	0.274	18.124	0.564	0.313
	MEDnerf	24.452	0.757	0.135	22.056	0.556	0.153
	FREEnerf	24.942	0.775	0.163	21.969	0.537	0.251
	ours	25.219	0.787	0.089	22.671	0.571	0.135
-5° to 5°	PIXELnerf	21.094	0.621	0.456	20.963	0.492	0.595
	GRAF	18.863	0.662	0.28	18.709	0.547	0.254
	MEDnerf	28.429	0.821	0.075	23.949	0.644	0.092
	FREEnerf	28.815	0.782	0.09	23.616	0.696	0.087
	ours	29.018	0.826	0.037	23.911	0.631	0.073

(Schwarz *et al* 2020), methods that fine-tune network parameters based on reference views (MEDnerf (Corona-Figueroa *et al* 2022) and FREEnerf (Yang *et al* 2023)) exhibit distinct advantages. MEDnerf (Corona-Figueroa *et al* 2022) achieves the best PSNR results in the chest region at −5° to 5°, but struggles to maintain performance away from the reference view.

4.2.2. Qualitative comparison

Figure 3 illustrates the visual results of ACnerf compared to other competitors in the 0 to 360° view. We also provide depth maps for each viewpoint. It can be observed that ACnerf maintains consistent stability across different viewpoints, and the utilization of the pose judgment network (ours) significantly enhances pose

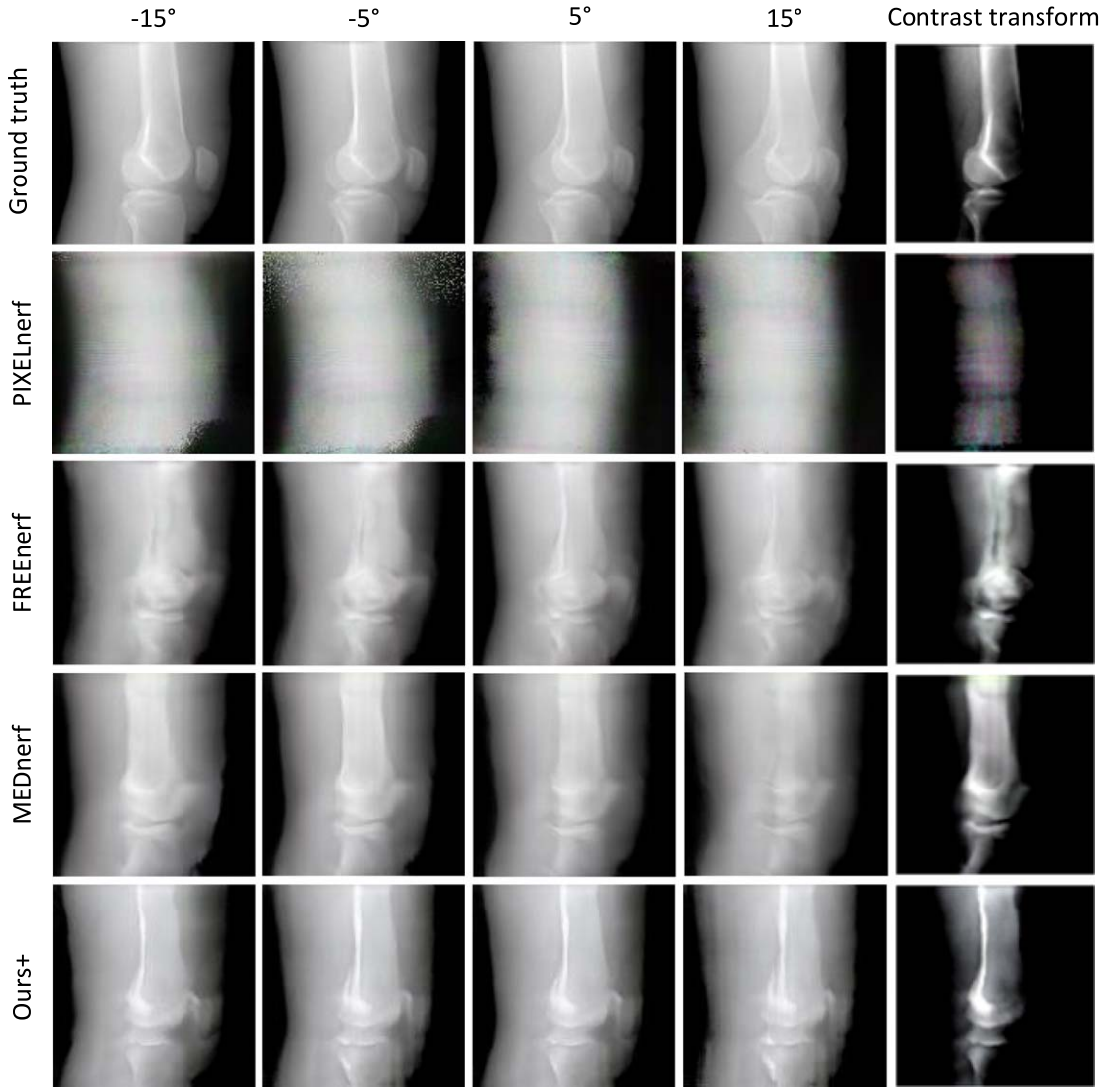

Figure 4. Visualization results of the −15° to 15° knee projection. The fifth column is the bone image extracted from the −15° projection using the contrast transform.

optimization. Although PIXELnerf's (Yu *et al* 2021b) PSNR is similar to ours, it notably falls behind in terms of LPIPS. This discrepancy stems from its projections having correct poses but lacking a substantial amount of high-frequency details, leading to a markedly inferior visual perception. From the depth maps, we can see its particles scattered on the surface of space. This issue is even more severe in FREEnerf (Yang *et al* 2023), where particles almost adhere to the outermost layer of space. GRAF (Schwarz *et al* 2020) generates the most visually appealing projections, but its pose does n0t match the ground truth, as analyzed in section 3.1. MEDnerf (Corona-Figueroa *et al* 2022) generates artifacts, and its pose still exhibits significant differences from the ground truth. The depth map shows that its spatial particles distort to positions without color in the ground truth.

In figure 4, we present the results of different methods rendered from −15° to 15°, with the 5th column showcasing the skeletal images extracted using contrast transform. In the context of localized projections, pose differences are almost indiscernible. However, through contrast enhancement, it can be observed that only ACnerf accurately restored the position and shape of the skeleton.

4.3. Ablation study

In this section, a detailed experimental justification is carried out for the three processes of ACnerf: (1) latent coded alignment projection; (2) pose judgment network; (3) fine-tuning of the generator parameters when masking in the frequency-domain. The baseline model used is GRAF (Schwarz *et al* 2020), performed on the knee dataset (Ali *et al* 2016).

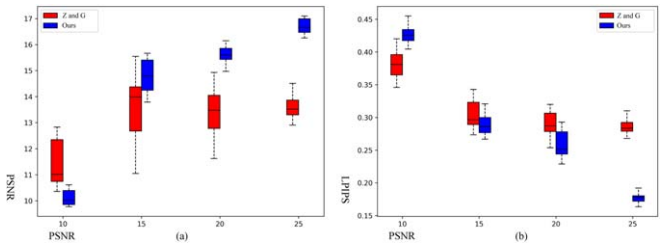

Figure 5. Comparison between the alignment of MEDnerf in section 3.1 (Z and G) and the alignment proposed in section 3.2 (Ours W/O PJ). The horizontal axes of both (a) and (b) denote the alignment values of the network with the reference projection. The vertical axes represent the average PSNR and LPIPS rendered by the network at the remaining poses, respectively.

Table 3. Comparison of different alignment methods. We compare the performance of the two alignment approaches reported in section 3.1 with the performance of our designed scheme in section 3.2 in terms of alignment peaks and alignment times as well as rendering results. The alignment peak is evaluated using PSNR, the alignment time in seconds, and the rendering results are evaluated using PSNR ↑, SSIM ↑, and LPIPS ↓. Optimal and suboptimal results are marked in red and blue, respectively.

	Only Z(GRAF)	Z and G(MEDnerf)	ours
Alignment peak values	24.114	32.547	27.085
Alignment times	57”	1’39”	1’8”
Rendered results	14.658/0.527/0.381	15.513/0.522/0.326	17.128/0.545/0.268

Table 4. The gain on the rendering results after correcting the pose. The second row (Original) is not aligned with the scheme presented in section 3.1, and the fifth row is the result obtained by correcting the pose after alignment. The rendering results are evaluated using PSNR ↑, SSIM ↑, and LPIPS ↓. The results in red in parentheses are the gains.

Method	GRAF			MEDnerf		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
Original	12.036	0.363	0.781	12.422	0.394	0.759
Pose Correction	14.007(1.971)	0.462(0.099)	0.416(-0.365)	14.885(2.463)	0.507(0.113)	0.414(-0.345)
Alignment	14.569	0.522	0.358	15.351	0.535	0.316
Pose Correction(aligned)	15.424(0.855)	0.538(0.016)	0.327(-0.031)	15.402(0.051)	0.541(0.006)	0.31(-0.006)

4.3.1. Use of latent codes

When aligning the pre-trained model with the reference projection, PSNR was employed as the assessment metric. Table 3 compares alignment peak values, alignment times, and rendered results among three approaches under a 360° projection: updating only latent codes, updating both the generator and latent codes, and utilizing latent codes to influence positional encoding. While updating the generator might lead to a higher alignment peak, our objective does not solely revolve around attaining alignment. Instead, our focus is on obtaining new viewpoint projections. Method 3.2 holds advantages regarding alignment time and rendering results.

We plotted box plots (figure 5) of the rendering results as the alignment outcomes grew. As the alignment results peak, updating the generator leads to divergence due to sacrificing the structure of unknown angles. In contrast, the alignment outcomes of method section 3.2 converge in the same direction as the rendering quality.

4.3.2. Pose correction

We have applied corrective measures using approach 3.3 to the adversarial loss-trained methods GRAF (Schwarz *et al* 2020) and MEDnerf (Corona-Figueroa *et al* 2022). Table 4 presents the rendering results for a 360° view. To enable the model to output results, we relaxed e to 0.08. Without aligning the model to the reference projection, the correction of pose improved both rendering results. However, upon aligning the model, only the method of keeping the generator parameters unchanged demonstrated a noticeable effect, as the inherent distortion in the projections caused by the network is irregular and uncontrollablewhich is also the motivation behind the design in section 3.2. the fine-tuning approach of section 3.4 is effective for generators aligned using all registration methods. In contrast to the levels reported in table 1, all methods exhibit varying degrees of decrease in the 360° projections.

Table 5. Rendering results after employing the frequency-domain regularization fine-tuned generator. PSNR ↑, SSIM ↑, and LPIPS ↓ were used for evaluation. Data in parentheses report differences from table 1 and table 2. Results labeled red in parentheses are their gains, and those marked blue are declines.

Method	0° to 360°			-15° to 15°		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
GRAF	14.126(-0.591)	0.490(-0.034)	0.397(0.037)	21.161(2.82)	0.686(0.061)	0.158(-0.116)
MEDnerf	14.475(-0.881)	0.535(-0.003)	0.356(0.045)	24.892(0.44)	0.773(0.016)	0.115(-0.020)
Ours	16.57(-1.238)	0.562(-0.022)	0.296(0.122)	25.219	0.787	0.089

5. Conclusion

With the advancement of NERF technology, many studies utilize it to recover medical data from a few images. Among them, inferring new views using only a single image poses is a significant challenges. We reveal that there is potential in two directions: (1) matching NERF models to medical images by more fully exploiting their generalization capability. (2) Refinement of the reasoning process to accommodate different rendering ranges. In this paper, we propose ACnerf, which reconstructs the projections of other pose from the projections of a single pose, with GRAF as the backbone. ACnerf utilizes latent codes to linearly influence positional encoding for image alignment, preventing changes in the generator's parameters from causing disruption to the 3D structure. During inference, a pose judgment network is employed to correct pose, optimizing the model's rendering views. For narrow-range projection rendering, we introduce a refinement technique involving the utilization of frequency-domain masks to fine-tune the generator. By adjusting alignment, inference, and rendering scopes, experimental results on knee and chest data with varying contrasts demonstrate that ACnerf outperforms the current state-of-the-art Radiance field approaches, effectively reducing artifacts and distortions.

Acknowledgments

This work was supported in part by the Science and Technology Commission of Shanghai Municipality under Grant 20DZ2254400 and 20DZ2261200; in part by Shanghai Municipal Science and Technology Major Project under Grant ZD2021CY001.

Data availability statement

All data that support the findings of this study are included within the article (and any supplementary information files).

ORCID iDs

Mengcheng Sun https://orcid.org/0009-0005-3533-6625
Yu Zhu https://orcid.org/0000-0003-1535-6520
Hangyu Li https://orcid.org/0000-0002-5793-298X

References

Ali A A, Shalhoub S S, Cyr A J, Fitzpatrick C K, Maletsky L P, Rullkoetter P J and Shelburne K B 2016 Validation of predicted patellofemoral mechanics in a finite element model of the healthy and cruciate-deficient knee *J. Biomech.* **49** 302–9

Barron J T, Mildenhall B, Tancik M, Hedman P, Martin-Brualla R and Srinivasan P P 2021 Mip-nerf: a multiscale representation for anti-aliasing neural radiance fields *Proc. of the IEEE/CVF Int. Conf. on Computer Vision* pp 5855–64

Barron J T, Mildenhall B, Verbin D, Srinivasan P P and Hedman P 2022 Mip-nerf 360: unbounded anti-aliased neural radiance fields *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition* pp 5470–9

Cen J, Zhou Z-W, Fang J, Shen W-M, Xie L, Jiang D, Zhang X and Tian Q 2023 Segment anything in 3D with nerfs arXiv:2304.12308

Chen A, Xu Z, Zhao F, Zhang X, Xiang F, Yu J and Su H 2021 Mvsnerf: fast generalizable radiance field reconstruction from multi-view stereo *Proc. of the IEEE/CVF Int. Conf. on Computer Vision* pp 14124–14

Chen Z, Funkhouser T, Hedman P and Tagliasacchi A 2023a Mobilenerf: exploiting the polygon rasterization pipeline for efficient neural field rendering on mobile architectures *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition* pp 16569–16578

Chen Z, Yang L, Lai J-H and Xie X 2023b Cunerf: cube-based neural radiance field for zero-shot medical image arbitrary-scale super resolution *Proc. of the IEEE/CVF Int. Conf. on Computer Vision* pp 21185–95

Cheng S, Chen Q, Zhang Q, Li M, Alike Y, Su K and Wen P 2023 Sdct-gan: reconstructing CT from biplanar x-rays with self-driven generative adversarial networks arXiv:2309.04960

Clark K et al 2013 The cancer imaging archive (TCIA): maintaining and operating a public information repository J. Digit. Imaging 26 1045–57
Corona-Figueroa A, Frawley J, Bond-Taylor S, Bethapudi S, Shum H P and Willcocks C G 2022 Mednerf: medical neural radiance fields for reconstructing 3D-aware CT-projections from a single x-ray 2022 44th Annual Int. Conf. of the IEEE Engineering in Medicine & Biology Society (EMBC) (IEEE) pp 3843–8
Fang Y, Mei L, Li C, Liu Y, Wang W, Cui Z and Shen D 2022 Snaf: sparse-view CBCT reconstruction with neural attenuation fields arXiv: arXiv:2211.17048
Guo Y, Chen K, Liang S, Liu Y-J, Bao H and Zhang J 2021 Ad-nerf: audio driven neural radiance fields for talking head synthesis Audio/CVF Int. Conf. on Computer Vision pp 5784–94
Huynh T, Gao Y, Kang J, Wang L, Zhang P, Lian J and Shen D 2015 Estimating CT image from MRI data using structured random forest and auto-context model IEEE Trans. Med. Imaging 35 174–83
Işık M, Rünz M, Georgopoulos M, Khakhulin T, Starck J, Agapito L and Nießner M 2023 Humanrf: high-fidelity neural radiance fields for humans in motion arXiv:2305.06356
Jain A, Mildenhall B, Barron J T, Abbeel P and Poole B 2022 Zero-shot text-guided object generation with dream fields Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition pp 867–76
Kasten Y, Doktofsky D and Kovler I 2020 Machine Learning for Medical Image Reconstruction: Third Int. Workshop, MLMIR, held in Conjunction with MICCAI 2020, Lima, Peru, Proc. 3 (Springer) pp 123–33
Kulhánek J, Derner E, Sattler T and Babuška R 2022 Viewformer: Nerf-free neural rendering from few images using transformers European Conf. on Computer Vision (Springer) pp 198–216
Li Y, Li K, Zhang C, Montoya J and Chen G-H 2019 Learning to reconstruct computed tomography images directly from sinogram data under a variety of data acquisition conditions IEEE Trans. Med. Imaging 38 2469–81
Lin Y, Luo Z, Zhao W and Li X 2023 Learning deep intensity field for extremely sparse-view CBCT reconstruction arXiv:2303.06681
Lindell D B, Martel J N and Wetzstein G 2021 Autoint: automatic integration for fast neural volume rendering Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition pp 14556–65
Liu L, Gu J, Zaw Lin K, Chua T-S and Theobalt C 2020 Neural sparse voxel fields Adv. Neural Inf. Process. Syst. 33 15651–63
Liu R, Wu R, Van Hoorick B, Tokmakov P, Zakharov S and Vondrick C 2023a Zero-1-to-3: zero-shot one image to 3D object Proc. of the IEEE/CVF Int. Conf. on Computer Vision pp 9298–309
Liu Y-L, Gao C, Meuleman A, Tseng H-Y, Saraf A, Kim C, Chuang Y-Y, Kopf J and Huang J-B 2023b Robust dynamic radiance fields Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition pp 13–23
Lo P et al 2012 Extraction of airways from CT (exact'09) IEEE Trans. Med. Imaging 31 2093–107
Martin-Brualla R, Radwan N, Sajjadi M S, Barron J T, Dosovitskiy A and Duckworth D 2021 Nerf in the wild: neural radiance fields for unconstrained photo collections Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition pp 7210–9
Max N 1995 Optical models for direct volume rendering IEEE Trans. Visual Comput. Graph. 1 99–108
Mescheder L, Geiger A and Nowozin S 2018 Which training methods for gans do actually converge? Int. Conf. on Machine Learning, PMLR pp 3481–90
Mildenhall B, Srinivasan P P, Tancik M, Barron J T, Ramamoorthi R and Ng R 2021 Nerf: representing scenes as neural radiance fields for view synthesis Commun. ACM 65 99–106
Müller T, Evans A, Schied C and Keller A 2022 Instant neural graphics primitives with a multiresolution hash encoding ACM Trans. Graph. (ToG) 41 1–15
Niemeyer M, Barron J T, Mildenhall B, Sajjadi M S, Geiger A and Radwan N 2022 Regnerf: regularizing neural radiance fields for view synthesis from sparse inputs Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition pp 5480–90
Pan X, Zhan X, Dai B, Lin D, Loy C C and Luo P 2021 Exploiting deep generative prior for versatile image restoration and manipulation IEEE Trans. Pattern Anal. Mach. Intell. 44 7474–89
Park K, Sinha U, Barron J T, Bouaziz S, Goldman D B, Seitz S M and Martin-Brualla R 2021 Nerfies: deformable neural radiance fields Proc. of the IEEE/CVF Int. Conf. on Computer Vision pp 5865–74
Poole B, Jain A, Barron J T and Mildenhall B 2022 Dreamfusion: text-to-3D using 2D diffusion arXiv:arXiv:2209.14988
Ranade S, Lassner C, Li K, Haene C, Chen S-C, Bazin J-C and Bouaziz S 2022 Ssdnerf: semantic soft decomposition of neural radiance fields arXiv:2212.03406
Roessle B, Müller N, Porzi L, Bulò S R, Kontschieder P and Nießner M 2023 Ganerf: leveraging discriminators to optimize neural radiance fields ACM Transactions on Graphics 42 1–14
Schwarz K, Liao Y, Niemeyer M and Geiger A 2020 Graf: generative radiance fields for 3D-aware image synthesis Adv. Neural Inf. Process. Syst. 33 0154–66
Shen L, Zhao W, Capaldi D, Pauly J and Xing L 2022 A geometry-informed deep learning framework for ultra-sparse 3d tomographic image reconstruction Comput. Biol. Med. 148 105710
Siddiqui Y, Porzi L, Bulò S R, Müller N, Nießner M, Dai A and Kontschieder P 2023 Panoptic lifting for 3D scene understanding with neural fields Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition pp 9043–52
Suetens P 2009 Fundamentals of medical imaging: visualization for diagnosis and therapy Available: CorpusID:57401532
Sun Y, Liu J, Xie M, Wohlberg B and Kamilov U S 2021 Coil: coordinate-based internal learning for tomographic imaging IEEE Trans. Comput. Imaging 7 1400–12
Trevithick A and Yang B 2021 Grf: learning a general radiance field for 3D scene representation and rendering Proc. of the IEEE/CVF Int. Conf. on Computer Vision pp 15162–72
Wang Q, Wang Z, Genova K, Srinivasan P P, Zhou H, Barron J T, Martin-Brualla R, Snavely N and Funkhouser T 2021 Ibrnet: learning multi-view image-based rendering Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition pp 4690–9
Wang T et al 2023 Rodin: a generative model for sculpting 3Ddigital avatars using diffusion Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition pp 4563–73
Wynn J and Turmukhambetov D 2023 Diffusionerf: regularizing neural radiance fields with denoising diffusion models Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition pp 4180–9
Wysocki M, Azampour M F, Eilers C, Busam B, Salehi M and Navab N 2023 Ultra-nerf: neural radiance fields for ultrasound imaging arXiv:2301.10520
Xie S, Huang W, Yang T, Wu D and Liu H 2020 Compressed sensing based image reconstruction with projection recovery for limited angle cone-beam CT imaging 42nd Annual Int. Conf. of the IEEE Engineering in Medicine & Biology Society (EMBC) (IEEE) pp 1307–10
Xie Y, Takikawa T, Saito S, Litany O, Yan S, Khan N, Tombari F, Tompkin J, Sitzmann V and Sridhar S 2022 Neural fields in visual computing and beyond Comput. Graphics Forum 41 641–76

Xu T and Harada T 2022 Deforming radiance fields with cages *European Conf. on Computer Vision* (Springer) pp 159–75
Yang J, Pavone M and Wang Y 2023 Freenerf: improving few-shot neural rendering with free frequency regularization *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition* pp 8254–63
Yu A, Li R, Tancik M, Li H, Ng R and Kanazawa A 2021a Plenoctrees for real-time rendering of neural radiance fields *Proc. of the IEEE/CVF Int. Conf. on Computer Vision* pp 5752–61
Yu A, Ye V, Tancik M and Kanazawa A 2021b pixelnerf: neural radiance fields from one or few images *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition* pp 4578–87
Zha R, Zhang Y and Li H 2022 Naf: neural attenuation fields for sparse-view CBCT reconstruction *Int. Conf. on Medical Image Computing and Computer-assisted Intervention* (Springer) pp 442–52
Zhang R, Isola P, Efros A A, Shechtman E and Wang O 2018 The unreasonable effectiveness of deep features as a perceptual metric *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition* pp 586–95