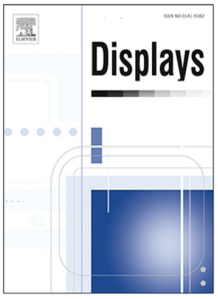
Journal Pre-proof



PGgraf: Pose-Guided generative radiance field for novel-views on X-ray

Hangyu Li, Moquan Liu, Nan Wang, Mengcheng Sun, Yu Zhu

PII: S0141-9382(26)00017-X
DOI: https://doi.org/10.1016/j.displa.2026.103354
Reference: DISPLA 103354

To appear in: *Displays*

Received date : 29 September 2025
Revised date : 25 December 2025
Accepted date : 13 January 2026

Please cite this article as: H. Li, M. Liu, N. Wang et al., PGgraf: Pose-Guided generative radiance field for novel-views on X-ray, *Displays* (2026), doi: https://doi.org/10.1016/j.displa.2026.103354.

This is a PDF of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability. This version will undergo additional copyediting, typesetting and review before it is published in its final form. As such, this version is no longer the Accepted Manuscript, but it is not yet the definitive Version of Record; we are providing this early version to give early visibility of the article. Please note that Elsevier's sharing policy for the Published Journal Article applies to this version, see: https://www.elsevier.com/about/policies-and-standards/sharing#4-published-journal-article. Please also note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2026 Published by Elsevier B.V.

Journal Pre-proof

PGgraf: Pose-Guided generative radiance field for novel-views on X-ray

Hangyu Li[a,*], Moquan Liu[a,*], Nan Wang[a,**], Mengcheng Sun[a], Yu Zhu[a,**]

[a]*School of Information Science and Engineering, East China University of Science and Technology, Shanghai, 200237, China*

Abstract

In clinical diagnosis, doctors usually judge the information by a few X-rays to avoid excessive ionizing radiation from harming the patient. The recent Neural Radiance Field (NERF) technology contemplates generating novel-views from a single X-ray to assist physicians in diagnosis. In this task, we consider two advantages of X-ray filming over natural images: (1) The medical equipment is fixed, and there is a standardized filming pose. (2) There is an apparent structural prior to X-rays of the same body part at the same pose. Based on such conditions, we propose a Pose-Guided generative radiance field (PGgraf) containing a generator and discriminator. In the training phase, the discriminator combines the image features with two kinds of pose information (ray direction set and camera angle) to guide the generator to synthesize X-rays consistent with the realistic view. In the generator, we design a Density Reconstruction Block (DRB). Unlike the original NERF, which directly estimates the particle density based on the particle positions, the DRB considers all the particle features sampled in a ray and integrally predicts the density of each particle. Experiments comparing qualitative-quantitative on two chest datasets and one knee dataset with state-of-the-art NERF schemes show that PGgraf has a clear advantage in inferring novel-views at different ranges. In the three ranges of 0° to 360°, -15° to 15°, and 75° to 105°, the Peak Signal-to-Noise Ratio (PSNR) improved by an average of 4.18 decibel, and the Learned Perceptual Image Patch Similarity (LPIPS)

*These authors contributed equally to this work.

**Corresponding author

Email addresses: wangnan@ecust.edu.cn (Nan Wang), zhuyu@ecust.edu.cn (Yu Zhu)

Journal Pre-proof

improved by an average of 50.7%.

Keywords: X-ray, Neural Radiance Field, Pose, Novel-view

1. Introduction

Computed tomography (CT) provides rich 3D data to assist physicians in diagnosis [1], based on scanning layers of a certain thickness with an X-ray beam to produce multiple 2D slices, which are later stacked into 3D data. This costs more in healthcare costs than a single X-ray image and also requires patients to be exposed to higher levels of ionizing radiation for more extended periods [2], increasing the risk of cancer. During the initial examination, physicians usually use their a priori medical knowledge to reason about 3D structures from a few X-ray images. Novel-view synthesis(NVS) for X-rays aims to reason about unknown viewpoint projections from existing X-ray projections. This scheme reduces the radiation dose and provides a more comprehensive view of the physician and downstream tasks such as CT reconstruction [3].

When the parameters of the inspection machine are known, 3D information can be reconstructed by building a mathematical model that matches its physical properties [4, 5]. Still, this condition also dramatically limits the model adaptation.

To overcome the performance limitations of the model due to unknown machine parameters, some studies have considered the use of deep learning techniques to restore 3D CT data while acquiring sparse viewpoint images [6, 7, 8]. However, these methods require not only 3D data supervision but also data labeling, which is very scarce in the fine-grained medical field. The neural radiance field(NERF) [9] can overcome the above limitations to some extent by using the 2D image itself as supervision, assuming that the space contains a certain number of light-emitting particles, and fitting an implicit function of the 3D model by neural network prediction of the particles' densities and colors, and then obtaining a novel-views through volume rendering. Nonetheless, the number of 2D images that can be obtained in a medical image novel-views reasoning task is tiny, e.g., the actual X-ray images taken are usually in the range of a few or one. This causes difficulties in obtaining more accurate 3D representations [10, 11] with primitive NERF techniques [9].

Journal Pre-proof

Journal Pre-proof

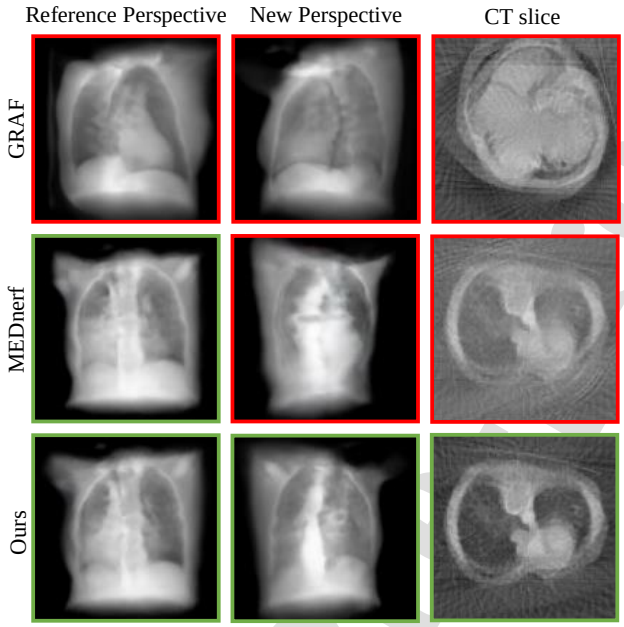


Figure 1: The first column of images is the model inference in the reference view, the second column is the inference away from the reference image, and the third column is the CT data slice reconstructed using the FBP [5] algorithm for the 72 views in the inference. Among them, images with severe distortions and viewpoint errors are marked by red boxes.

In the NVS task for natural images, many studies have improved NERFs regarding constraints on sparse-views and generalization capabilities to enhance the quality of novel-views when there are only a small number of image references [11, 12, 13, 14, 15, 16, 17, 18]. In medical imaging tasks, MEDnerf [19] draws on Generative Radiance Field (GRAF) [18] technology to enable continuous novel-view generation with only a single X-ray as a reference. This method provides ideas for novel-view reconstruction of medical images in sparse-view. However Generating complete 3D projections based on a single image relies heavily on a priori knowledge [11], and there are some differences between medical and natural scenes in this regard. Current, the task still presents significant challenges: On the one hand, the generated viewpoints do not correspond to the realistic viewpoints (Fig.1(GRAF)), and on the other hand, the 3D structure is destroyed when fine-tuning the model with reference to X-rays (Fig.1(MEDnerf)).

In this paper, we build on the above work by considering two advantages

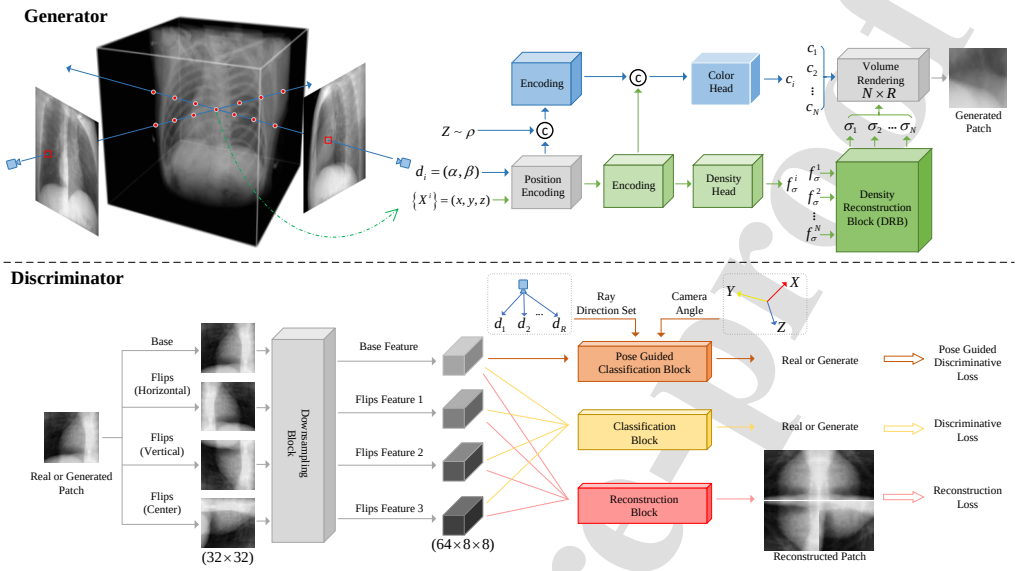


Figure 2: General structure diagram of PGgraf. In the generator, the color branch is marked blue, and the density branch is marked green. In the discriminator, the reconstruction branch is marked red, the discriminate branch is marked yellow, and the pose-guided discriminate branch is marked orange.

of medical image tasks over natural images:(1) The imaged poses are known due to the immobilization of the detection device. (2) Samples from different patients in the same pose have similarities. Combining such advantages, we propose a Pose-Guided generative radiance field named PGgraf, as shown in Fig.2. PGgraf uses Generative Radiance Field (GRAF) [18] as a backbone to fuse the pose information with the image information in the discriminator, which together serves as the basis for the discrimination of image authenticity, prompting the novel-views of model inference to be consistent with the reality of the medical scenario. In order to fully utilize the feature prior of X-ray images of the same class, we designed a Density Reconstruction Block(DRB) in the generator of PGgraf. We simulated paired X-ray images on each of the three CT datasets (containing knee and chest) to validate the ability of the proposed model to synthesize novel-views when referring to a single X-ray. Comprehensive experiments on synthesizing new viewpoints in different angular ranges show that PGgraf can generate viewpoint-consistent X-rays stably, with a clear performance advantage over state-of-the-art NERF models. The main contributions of this paper are summarized below:

Journal Pre-proof

- We propose a Pose-Guided generative radiance field (PGgraf), which employs a Pose guidance strategy to fuse pose and image information during training. In the inference phase, the PGgraf can generate X-rays that are consistent with realistic angles, thus enabling the radiance field to correspond to the medical case.

- Considering that samples of the same pose from different patients have similarities, we designed a Density Reconstruction Block (DRB) in the generator of PGgraf. DRB can make a comprehensive determination of particle density based on particle characteristics at different locations. The method allows efficient model fine-tuning in reference to a single X-ray using a priori information from the same site.

- Qualitative-quantitative experiments on two chest datasets and one knee dataset show that PGgraf outperforms the state-of-the-art NERF methods in novel-news inferring tasks across a range of angles. Across the three angle ranges of 0° to 360°, -15° to 15°, and 75° to 105°, PSNR improved by an average of 4.18Db, SSIM improved by an average of 0.074, and LPIPS improved by an average of 50.7%.

2. Related work

Neural Radiance Field (NERF) techniques have been developing rapidly in recent years, and some preliminary results have appeared in the medical field. In this section, we first present research on sparse-views NERF relevant to the task of this paper. After that, the applications of NERF in medicine are reported. Finally, in more detail, we review the NERF technique and the backbone (GRAF) used in this paper.

2.1. NERF with sparse-views

When only a small number of views (more than five) are available as references, synthesizing novel-views requires more a priori information as a reference. Considering the feature maps [11, 12] inferred by the pre-trained model or the depth information of the reference views [13] as additional inputs to the neural network can enhance the quality of the novel-vews. Meanwhile, in some specific scenarios where geometric priors exist [14, 15, 16], corresponding loss functions can be designed to govern the generation of novel-views. However, when X-rays are taken in the clinic, only one to three

Journal Pre-proof

images are usually taken for the initial examination, considering the damage caused by radiation to the human body. This dramatically limits the ability of the above model. When using a single reference image to generate novel-views, some studies have considered intervening pre-trained diffusion models [20, 21] into the training of NERF, often accompanied by expensive computational costs. Also, the a priori knowledge of large diffusion models is mainly derived from natural images, which makes it difficult to achieve precise control over the views of medical images—specifically, the models require the professional judgment of human physicians when determining the views of generated medical images. The use of adversarial training [17, 18] in a defined class of images allows the NERF model to capture the structure of the manifold, and then edit the output of the model through tricks such as latent code to obtain the novel-views of the target scene. Such methods have achieved excellent results with natural images. Still, when considering medical images, which are extremely demanding not only in terms of perception but also in terms of accuracy, targeted improvements need to be made.

2.2. Application of NERF in medical image

Based on the success of NERF in synthesizing novel-views of natural images, medical data reconstruction is beginning to draw on such ideas. MRI data [22] and CBCT data [23, 24] can be reconstructed directly in 3D by imaging principles or 3D data supervision during training with multiple views for reference. These tasks also require at least ten images when considering sparse-views. Currently, in extremely sparse cases, reasoning about novel-views rather than the complete 3D data is a more viable option. MEDnerf [19] proposed a challenging task based on the clinical acquisition of X-rays, using a single X-ray as a reference to generate a corresponding 0° to 360° views. It uses the Generative Radiance Field [18] (GRAF) as a framework to fine-tune the model in reference to a single X-ray, and then renders the novel-views. Among other things, GRAF uses random poses as inputs during training to ensure generalization capabilities and supervises only the authenticity of the generated views. This leads to the inference that the novel-views do not correspond to the realistic ones, making it difficult to meet the accuracy requirements of medical images.

2.3. NERF overview

The NERF assumes that the space contains light-emitting particles with density, and estimates their density (σ) and color (c) using the particle's

Journal Pre-proof

position $X = (x, y, z)$ and direction angle (d) as inputs to the neural network. Sampling N particles in space along the ray $r(t) = o + td$ allows for an approximation of the pixel color $C(r)$ in that direction based on the volume rendering [25]:

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

In practice, the inputs to the neural network are encoded at high frequencies in a process that can be described as follows:

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

The high-frequency encoded particle position information $\gamma(x_k)$ and direction information $\gamma(d)$ are used as inputs to the neural network to estimate its density and color:

$$(c_k, \sigma_k) = F_{MLP}(\gamma(x_k), \gamma(d)) \tag{3}$$

In the original NERF [9] model, the pixel values of the reference image are used as supervision and the reconstruction loss is constructed to compute the difference between the rendered pixels and the real pixels to train the neural network. This results in strict network inference of particle densities, and GRAF introduces the latent code Z into the radiance field, allowing particles at the same location to be inferred with different densities and colors. Meanwhile, the loss function is an adversarial loss constructed by a discriminator [26], which determines the truthfulness of the patches obtained from random poses.The process is described below:

$$\begin{aligned} L(\theta, \phi) = &\mathbb{E}_{Z \sim N(\vec{0}, I)}[f(D_\phi(G_\theta(Z, \xi, v)))] \\ &+ \mathbb{E}(f(-D_\phi(I) + \lambda \left|\left|\nabla D_\phi(I)\right|\right|^2)) \\ &, f(t) = -log(1 + exp(-t)) \end{aligned} \tag{4}$$

Where Z is the 128-dimensional latent encoding sampled in a Gaussian distribution, ξ and v are the set of particle positions and observation directions required for rendering. MEDnerf considers its application to medical images by training the model on X-rays with different contrasts. Only one X-ray was provided as a reference for testing, and the parameters of the entire network were fine-tuned [27] to ensure that the rendered X-rays at the

Journal Pre-proof

same angle were consistent with the reference. This scheme corresponds the model-generated view to the realistic view to some extent, but it is highly susceptible to overfitting, causing distortion and blurring of the novel-views.

3. Method

In this section, we first introduce the proposed network structure of the Pose-Guided generative radiance field (PGgraf). Then, we introduce two schemes of pose-guided radiance field in detail. Finally, we describe how the designed Density Reconstruction Block (DRB) acts and the fine-tuning and inference strategy when referring to a single X-ray.

3.1. Network structure

We designed PGgraf with reference to the overall structure of GRAF [18], containing a generator (Neural Radiance Field) and a discriminator, as shown in Fig.2. Among them, the generator (Fig.2.generator) is similar to the one introduced in section 2.3, taking as input the position information $X = (x, y, z)$ of the particles in space and the direction information d of the observation. The two types of information are used as inputs to the color branch (upper half) and the density branch (lower half) after high-frequency encoding, respectively.

In the color branch, the direction information is combined with the randomly sampled latent code Z, which is then passed through an encoder to obtain the direction feature. This is then combined with the obtained positional features in the density branch, and finally, the color c is predicted by the color head. In the density branch, we design the Density head not to predict the density directly, but to output the features f_σ. The features of N particles sampled in a ray are used as inputs to the DRB, which in turn yields the density of N particles. DRB can be integrated based on the features of neighboring particles, and is more capable of learning the structural prior of the image during training than predicting each density in isolation. Finally, the particle information (c, d) on the R rays is volume rendered to get the predicted patches.

The discriminator (Fig.2.discriminator) consists of three branches: the reconstruction branch, the discriminator branch, and the pose-guided discriminator branch. In the reconstruction branch, the input patch is first passed through the downsampling module to get the image features. After

Journal Pre-proof

that, it is restored to the input patch by the reconstruction block, and the Reconstruction Loss L_R expression is as follows:

$$L_R = MSE(P_{input}{}', P_R) \tag{5}$$

Where MSE denotes the mean square error, and $P_{input}{}'$ and P_R denote the stitching of the patchs of the four branches and their corresponding reconstructed patchs, respectively. This process is similar to Auto-Encoder; its primary role is to regularize the parameters of the downsampling block during the training process.

The discriminative branch predicts the truthfulness of the input patches based on the downsampled features using the classification block; we draw on the least squares loss method [28] , and the expression for the Discriminative Loss L_D is as follows:

$$\begin{aligned} L_D = &\frac{1}{2}E_{P\sim real}[D(P)-1]^2 \\ &+\frac{1}{2}E_{P'\sim generated}[D(P')-A]^2 \end{aligned} \tag{6}$$

Where P denotes the real patch sampled from the training set, P' denotes the predicted patch inferred by the generator, and D denotes the arithmetic of the discriminate branch. When updating the discriminator parameters, $A = 0$, and when updating the generator parameters, $A = 1$. In the above two branches, we flip the patches horizontally, vertically, and centrally, for a total of four images to training models.

In the pose-guided discriminative branch, only a base patch is used for training; features are fused with pose information, and finally, authenticity is predicted by a classification block. The specific expression for the Pose Guided Discriminative Loss L_{pose} is as follows:

$$\begin{aligned} L_{pose} = &\frac{1}{2}E_{P_{base}\sim real}[[D(P_{base}, Pose_r)-1]^2 \\ &+[D(P_{base}, Pose_c)-1]^2] \\ &+\frac{1}{2}E_{P'_{base}\sim generated}[[D(P'_{base}, Pose_r)-A]^2 \\ &+[D(P'_{base}, Pose_c)-A]^2] \end{aligned} \tag{7}$$

Where P_{base} and P'_{base} denote the real and predicted base patch, respectively. $Pose_r$ and $Pose_c$ denote the ray direction and camera angle corresponding to the pose information at the time of the shooting, respectively, which we describe in detail in the next section.

Journal Pre-proof

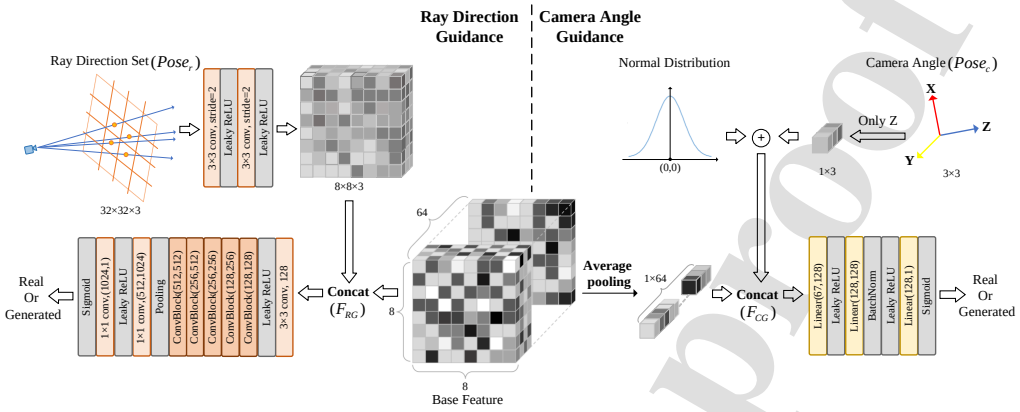


Figure 3: Pose Guided Classification Block. The left half is the fusion process for the set of ray directions, and the right half is the fusion process for the camera angle. The output of both parts is a value between 0 and 1, indicating the likelihood that the patch is real.

During training, the total loss function of the model is expressed as follows:

$$L_{total} = \lambda_1 L_R + \lambda_2 L_D + \lambda_3 L_{pose} \tag{8}$$

Where L_D and L_{pose} update the parameters of the discriminator and generator with the same update scheme, while L_R updates only the discriminator. Where parameters λ_1, λ_2 and λ_3 are set is reported in the experiment section. Detailed configurations of the hyperparameters λ_1, λ_2, and λ_3 are reported in section 4.2. The detailed inference process and fine-tuning strategy we report in section 3.3.

3.2. Pose guidance strategy

Unlike the random placement of objects in natural images, the human body's relative angle (pose) to the machine is known when taking an X-ray. To take advantage of this, we introduce pose information into the discriminator to guide the generator to infer X-rays consistent with a realistic view. The process exists in two ways of pose information introduction, as shown in Fig.3. The pixels of an X-ray are determined by the direction of the rays in the radiance field, and the set of ray directions $Pose_r$ of the patch corresponds to the pose at the time of the shot. The left part of Fig.3 demonstrates the scheme of ray direction set as pose information; after the pose information has been subsampled, it is spliced with the image features output from the

Journal Pre-proof

attention module, which serves as the judgment information, the process is described below:

$$F_{RG} = Concat(Subsampled\underbrace{(d_1, d_2, ...)}_{32\times32}, F_{patch}) \tag{9}$$

where F_{patch} denotes the Base Feature after downsampling in Fig.2. F_{RG} indicates the guiding features (judgmental information) after the fusion of F_{patch} with ray direction information.

The right side of Fig.3 shows the fusion strategy for the camera's pose information. Relative to the world coordinates, the direction of rotation of the shot (Z-axis) is the pose of the shot. The patch features are pooled evenly by channel and spliced with the pose information as the judgment information of the patch authenticity. Since there is a limit to the number of views that can be provided during actual training, we added Gaussian noise to the pose information to prevent the generator from overfitting. The process is as follows:

$$F_{CG} = Concat\left(N(\vec{0}, I) \oplus Pose_c, Avgpool\left(F_{patch}\right)\right) \tag{10}$$

Where $Pose_c$ denotes the camera angle, which is summed with the noise sampled from a standard Gaussian distribution. F_{CG} indicates the guiding features obtained by pooling F_{patch} by channel and fusing it with the camera pose information.

Both strategies were used simultaneously during training, and we report the gain of each strategy in the section 4.3.

3.3. Inference processes and fine-tuning strategies

After model training on a class-specific X-ray dataset (chest or knee) is complete, only the generator is used in the inference phase. Changing the input to the generator, a continuous arbitrary viewpoint image can be obtained by the volume rendering(Equation (1)). At the same pose, different latent code Z can obtain various shapes of X-ray, as shown in 4(a). However, using only Z as a parameter for fine-tuning has limited expressive power, and it is difficult to make the X-ray generated from the same pose consistent with the reference X-ray.

In order to align more efficiently with the reference X-ray, we designed the Density Reconstruction Bloc(DRB), as shown in Fig.4(b). In the generator,

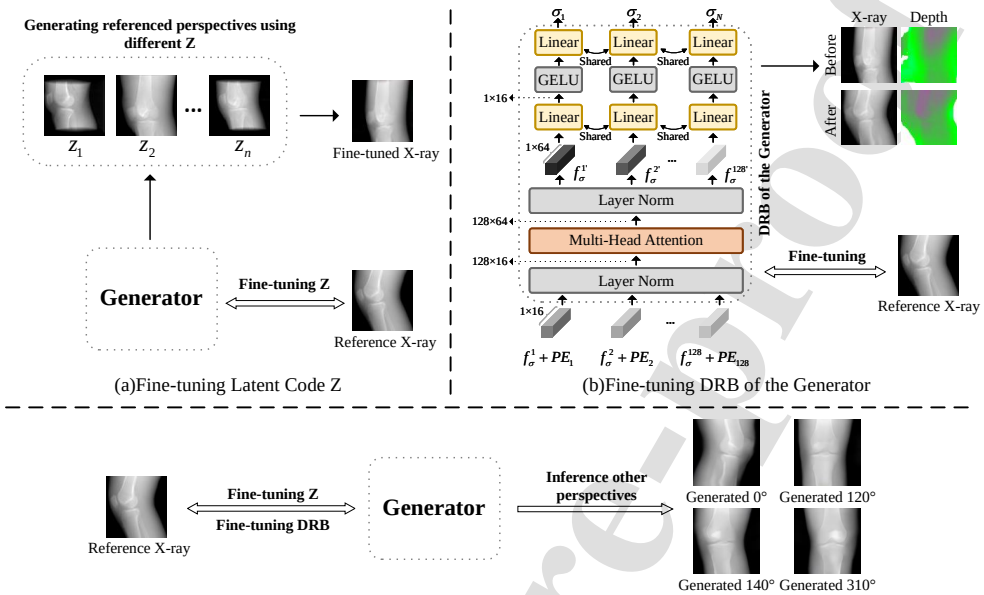


Figure 4: Fine-tuning strategies and inference. Among them, part (a) shows the changes in the X-ray by changing the input latent code Z. Part (b) shows the detailed structure of the DRB and its role in fine-tuning. Part (c) shows the fine-tuning and inference process.

the DRB combines the density feature f_σ of the ray-sampled 128 points with the sequence information PE [29] as input. The 128 density features f'_σ considered in the synthesis are obtained after a multi-head attention block, and each feature is individually subjected to a regression network to obtain the particle density σ. Compared to the original NERF, this process adds a small computational cost but allows particle information to interact. When referring to a single X-ray, the module can be fine-tuned to approximate the density distribution of the target 3D structure, thus realizing the alignment of the X-ray at the same pose. The overall fine-tuning process is as follows:

$$\theta_{DRB}, Z = \arg\min_{\theta_{DRB}, Z} MSE(I, I') \tag{11}$$

where I and I' denote the reference and generated X-rays of the same pose, respectively.

After the model is aligned with the reference X-ray, the novel-view is obtained by modifying the inputs to the generator, as shown in Fig.4(c). In the

Journal Pre-proof

experimental section, we report the effect of different fine-tuning strategies on the novel-views.

4. experiment

4.1. Dataset and metrics

Acquisition of multiple X-rays with different poses would expose the patient to intolerable radiation, and we refer to the scheme of MEDnerf [19] to simulate X-rays with varying views from CT data. The CT data included: (1) TCIA [30] dataset containing thoracic data from 20 patients with different degrees of analogy. (2) NATURAL KNEE DATA [31], containing CT data from 6 raw knees. (3) COVID-DS36, a dataset provided by the partner hospital, contains CT data of 36 patients with an age distribution of 6 to 66 years old, with distinct imaging characteristics. During the simulation, rotating horizontally around the CT data center, one X-ray was generated every 5°, and 72 X-rays with a resolution of 128×128 were obtained for each CT. Therefore, this work does not involve experimental procedures on human subjects or animals. We randomly selected 75%-84% of the complete data as the training set (15 TCIA, 5 NATURAL KNEE DATA, and 30 COVID-DS36), and the remaining patient data were used for testing. For testing, X-rays from one randomized view are used as a reference, and the remaining 71 views are used to evaluate the model performance.

We evaluate the quality of the inferred novel-views based on three visual metrics: (1)Peak Signal-to-Noise Ratio (PSNR), which measures the pixel difference; the higher the value, the smaller the difference. (2)Structural similarity (SSIM) measures pixel-level correlation, with values closer to 1 being more similar. (3)LPIPS [32] measures the difference in the perception of an image by a pre-trained model; the lower the value, the smaller the difference. All experiments are based on the pytorch framework and run on a single NVIDIA RTX A6000 GPU.

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

Among the compared methods, two models with excellent generalization capabilities introduced in Section 2.3, GRAF [18] and MEDnerf [19], are included. Two models applied to sparse-views reconstruction, PIXnerf [11] and FREEnerf [16], are also included. The PIXnerf takes the a priori information of the pre-trained network as part of the input to obtain generalization capability and support the reconstruction of a single image. FREEnerf, on

Journal Pre-proof

Table 1: Detailed structure of Reconstruction Block

Reconstruction Block	
UpBlock(in,out)	Upsample(2,'nearest')
	Conv2d(in,out,3×3,1)
	BatchNorm2d,GLU
Sequential	UpBlock(64,128)
	UpBlock(128,128)
	Conv2d(128,3,3×3,1)
Dimension	input=64×8×8
	output=3×32×32

the other hand, considers regularized inputs for positional coding to optimize the training process and does not have generalization capability. For testing, we provide 18 views to FREEnerf for training, demonstrating comparable references to other models. The Encoding, Head, Downsampling Block, Classification Block of our proposed model PGgraf are consistent with the GRAF [18] report. The detailed structure of the designed Reconstruction Block is shown in Table 1. The coding dimensions for position and direction were uniformly 63 and 27 for all models, respectively. During training, Adam was used as the optimizer for 100,000 iterations, with a batch size of 16 and a learning rate ranging from 10^{-4} cosine annealing to 10^{-6}. When the model(PGgraf, GRAF, MEDnerf) needs to align the reference image, we consistently use PSNR as the metric, and the alignment is complete when the increase in PSNR is less than 0.5% in every 50 iterations. The loss function ratios reported in Section 3.1 $\lambda_1 = \lambda_2 = 1$, with λ_3 initially set to 10, decreasing by 0.25 per 1250 iterations and decreasing to 1 to stop.

4.2.1. Quantitative comparison

Table 2 reports the average quality of the 0° to 360° views generated by the different methods. Depending on whether a DRB is used or not, we report two versions of PGgraf. On all three datasets, both PGgraf schemes offer significant advantages. On the metric LPIPS, which is highly correlated with vision, there was an average improvement of nearly 49% compared to the baseline model we used, GRAF. Based on the observation of Ours w/o DRB results, the introduction of pose information significantly gains the model's performance.

Journal Pre-proof

Table 2: Comparison of the performance of different methods for generating 0° to 360°.The best and second-best results are labeled in red and blue, respectively.

dataset	Evaluation	PIXELnerf	GRAF	MEDnerf	FREEnerf	our w/o DRB	ours
NATURAL KNEE[31]	PSNR↑	17.059	14.717	15.356	12.641	17.519	17.721
	SSIM↑	0.529	0.524	0.538	0.308	0.582	0.589
	LPIPS↓	0.469	0.360	0.311	0.434	0.171	0.168
TCIA[30]	PSNR↑	16.874	15.493	15.524	13.173	18.485	18.546
	SSIM↑	0.335	0.328	0.339	0.288	0.463	0.459
	LPIPS↓	0.432	0.337	0.326	0.621	0.186	0.180
COVID-DS36	PSNR↑	18.593	14.940	16.603	13.855	20.887	21.206
	SSIM↑	0.417	0.431	0.440	0.413	0.503	0.511
	LPIPS↓	0.442	0.359	0.327	0.447	0.195	0.192

In general, the model can generate more accurate results in its neighborhood based on the reference image. Table 3 reports the generation quality from -15° to 15° from the reference views. Only FREEnerf, which references 18 views, slightly outperforms our model on both datasets, which side-steps the competitiveness of PGgraf in sparse-views.

Table 3: Comparison of the performance of different methods for generating −15° to 15°.The best and second-best results are labeled in red and blue, respectively.

dataset	Evaluation	PIXELnerf	GRAF	MEDnerf	FREEnerf	our w/o DRB	ours
NATURAL KNEE[31]	PSNR↑	19.469	18.341	24.452	24.942	24.638	24.930
	SSIM↑	0.569	0.625	0.757	0.775	0.771	0.779
	LPIPS↓	0.542	0.274	0.135	0.163	0.133	0.131
TCIA[30]	PSNR↑	18.551	18.124	22.056	21.969	22.047	22.084
	SSIM↑	0.488	0.564	0.556	0.537	0.572	0.575
	LPIPS↓	0.582	0.313	0.153	0.251	0.147	0.149
COVID-DS36	PSNR↑	20.148	20.510	23.487	26.195	25.615	25.637
	SSIM↑	0.424	0.458	0.465	0.516	0.530	0.537
	LPIPS↓	0.594	0.486	0.186	0.190	0.182	0.180

The reference image provides only a tiny amount of information in the direction perpendicular to the viewing angle; in contrast, accurately predicted X-rays in the vertical viewing angle provide more structural information. Table 4 reports the generation quality from 75° to 105° from the reference viewing angle. In this range, the DRB produces gains in all metrics. Among models of the same type that require fine-tuning to align the reference (PGgraf, GRAF, MEDnerf), PGgraf has the lowest performance degradation compared to the other ranges. While FREEnerf has a signifi-

Journal Pre-proof

cant performance drop in this range due to a lack of a priori information.

Table 4: Comparison of the performance of different methods for generating 75° to 105°.The best and second-best results are labeled in red and blue, respectively.

dataset	Evaluation	PIXELnerf	GRAF	MEDnerf	FREEnerf	our w/o DRB	ours
NATURAL KNEE[31]	PSNR↑	16.215	14.428	13.485	12.181	16.270	16.276
	SSIM↑	0.516	0.518	0.509	0.306	0.557	0.562
	LPIPS↓	0.504	0.373	0.367	0.452	0.246	0.243
TCIA[30]	PSNR↑	15.752	15.409	14.127	12.864	17.735	17.769
	SSIM↑	0.450	0.513	0.534	0.511	0.554	0.561
	LPIPS↓	0.438	0.340	0.339	0.626	0.216	0.211
COVID-DS36	PSNR↑	18.348	14.507	14.974	12.894	19.914	19.926
	SSIM↑	0.414	0.426	0.428	0.301	0.478	0.484
	LPIPS↓	0.636	0.522	0.371	0.479	0.206	0.202

4.2.2. Qualitative comparison

Fig.5 illustrates the different methods' 0° to 360° views generated using 0° as a uniform reference. Columns 1,3,5 show the inferred novel-views, and columns 2,4,6 perform the error maps. We can observe that in high-frequency X-ray regions, such as the heart area, PGgraf has less error than other methods. And GRAF does not generate the correct view, which motivates our design. The quantitative data show that PIXnerf performs very poorly on the LPIPS, and the figure also shows that it does not generate valid high-frequency information.

Fig.6 shows the performance of the different methods when perpendicular to the reference projection. FREEnerf can barely generate discernible X-rays, and the depth maps of the particles (columns 2,4) show that their spatial distribution is not uniform. MEDnerf also did not obtain accurate X-rays, similar to those analyzed in 2.3, where the 3D structure was destroyed during the fine-tuning process, and the particles were mostly stuck to the surface. The two PGgraf schemes had a more homogeneous distribution of the particles and were farther away from our observing surface.

Although the generalized models (PGgraf, GARF, MEDnerf) are obtained by training on the full range of angles, the variation of their generated angles is not necessarily uniform. Fig.7 shows the comparison of different methods for generating continuous views containing contrast transformation, L-K optical flow maps, and 3D CT slices reconstructed from 72 views using

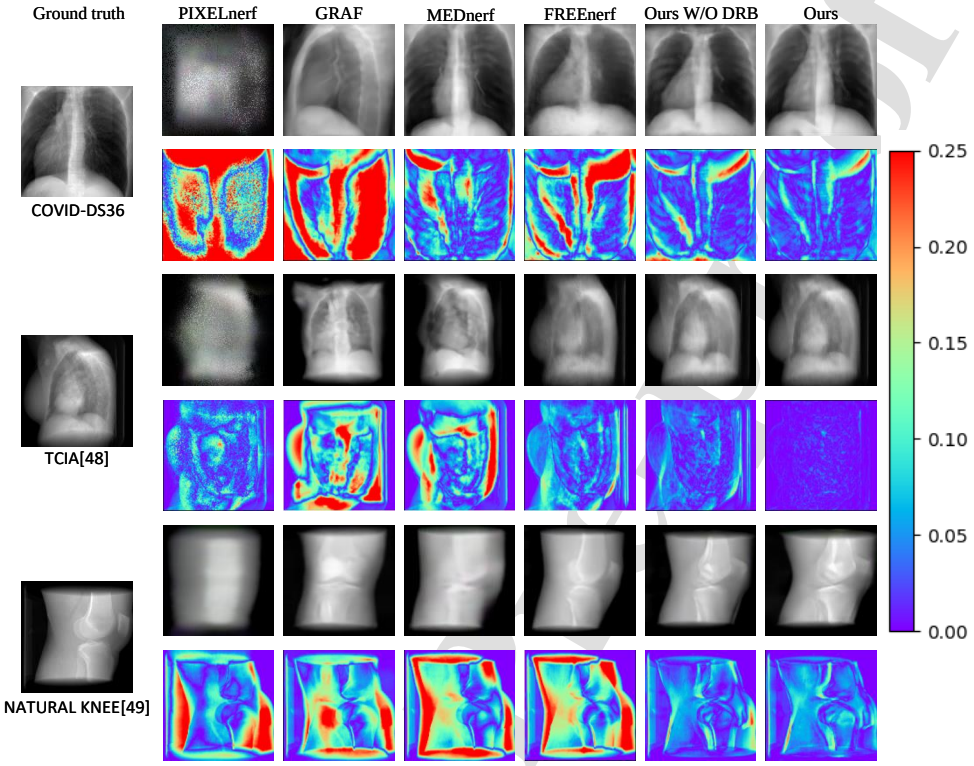


Figure 5: 0° to 360° generated views. Where the first, third, and fifth rows generate angles of 45°, 150° and 210° respectively. The second, fourth, and sixth rows are heat maps of pixel differences.

the FBP method. By transforming the view contrast, PGgraf shows skeletal images similar to Ground truth, whereas GRAF generates erroneous structures and MEDnerf has artifacts. The optical flow map shows that our corner point locations and variations are more consistent with Ground truth. Although the new views reasoned in this task to reconstruct 3D CT data were difficult, our method kept the slice angles correct. MEDnerf and GRAF showed strong artifacts and angular errors, respectively.

4.3. Ablation study

This section demonstrates a detailed experimental analysis of the pose information fusion strategy proposed in Section 3.2 and the fine-tuning strategy proposed in Section 3.3. The datasets for the experiments are COVID-DS36 to verify the model's performance in generating 0° to 360° views.

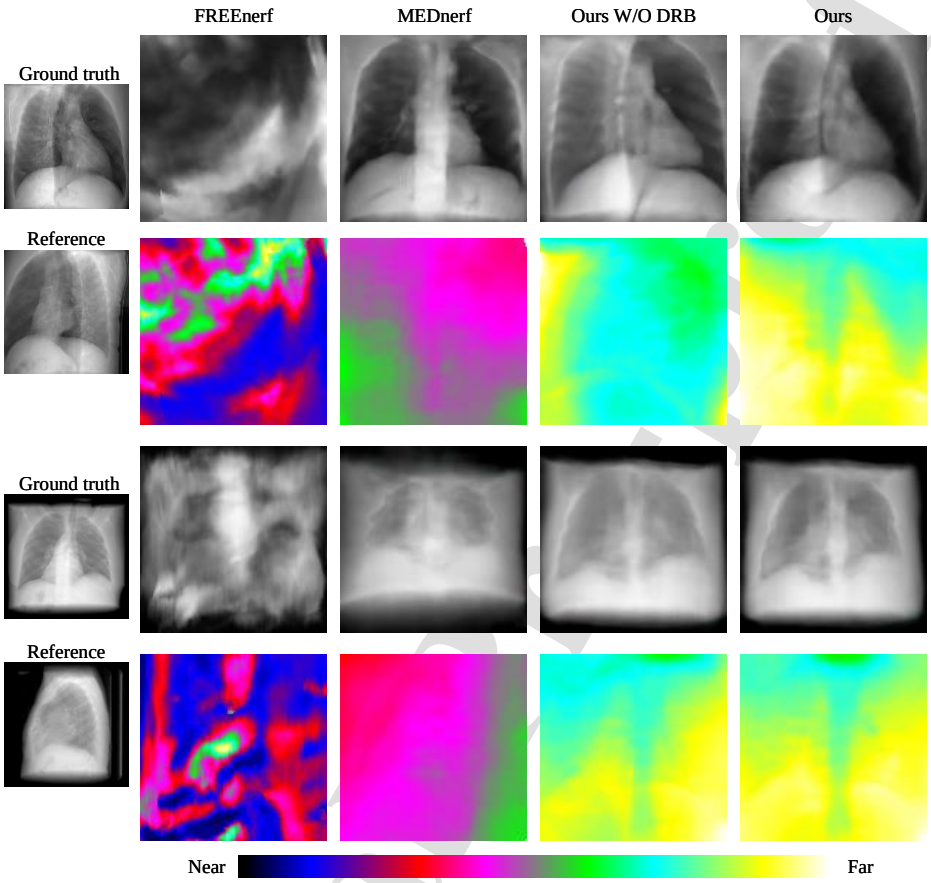


Figure 6: Generated views perpendicular to the reference x-ray.where the second, fourth rows are depth maps of the particle distribution.

Table 5: Impact of different pose guidance strategies on model performance

Evaluation	no pose	ray	camera	combinatorial
PSNR↑	15.165	18.924	20.634	20.887
SSIM↑	0.462	0.494	0.496	0.503
LPIPS↓	0.338	0.213	0.201	0.195

4.3.1. Pose information fusion

Section 3.2 details two ways of fusing pose information in the discriminator: the ray direction set and the camera pose. Although both types of

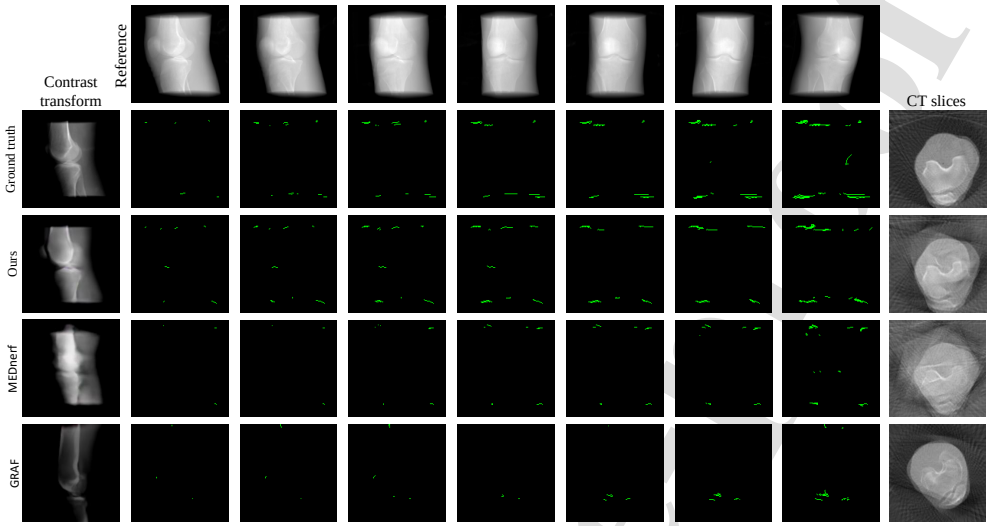


Figure 7: Comparison of Perspective Changes in Different Methods. The first row is a real X-ray from the observation views. The first column is skeletal images extracted by contrast transformation. The second to seventh columns are the optical flow maps generated by the angular variation. The last column is the 50th layer slice of the 3D CT data recovered by the FBP method using 72 views.

information correspond to views, the amount of information varies, making the results different. Table 5 shows the model's performance trained with different pose information (without the DRB). When trained without pose information, the performance is close to that of the GRAF reported in Table 2. The model performance improves substantially after using any of the pose guidance strategies, outperforming all the competitors in Table 2. Compared to the camera pose guidance, the ray direction information is more dense, increasing the difficulty of training, and its performance is slightly lower.

Table 6: Impact of reference on model performance. Calculated indicator is LPIPS↓.

	GRAF	MEDnerf	Ours
no reference	0.476	0.453	0.247
reference	0.359	0.327	0.195

Combining the two approaches further improves the performance of the model. Table 6 shows the LPIPS performance without reference to X-rays

Journal Pre-proof

after the model has been trained, and the comparison scenario contains the two baseline models we used (GRAF and MEDnerf). After the same training in generalization skills, combining pose information has an advantage in perception.

Table 7: Performance of different fine-tuning schemes

	PSNR↑	SSIM↑	LPIPS↓
DRB	19.683	0.470	0.219
z	20.891(+1.208)	0.505(+0.035)	0.198(-0.021)
all	21.206(+0.315)	0.511(+0.006)	0.192(-0.006)

4.3.2. Fine-tuning strategy

In Section 3.3, we introduced fine-tuning the DRB and the latent code Z to align the reference X-rays. Table 7 shows the performance obtained with the different approaches, where the models all include the DRB structure and are trained in the optimal way introduced in 4.3.1. Compared to the results reported in Table 6(no reference), both approaches produce gains. Also, fine-tuning the DRB is modifying a small portion of the parameters of the network itself, while the latent code Z is an input to the network, and the two approaches do not conflict. The maximum gain in performance is obtained when both schemes are used simultaneously.

Table 8: FBP algorithm to reconstruct 3D CT data from 72 views

Evaluation	GRAF	MEDnerf	Ours(W/O DRB)	Ours(w/ DRB)
PSNR↑	13.658	14.706	17.133	17.721
SSIM↓	0.421	0.458	0.492	0.509

Table 8 demonstrates the effect of whether the DRB module is employed on the sparse 3D CT reconstruction, where 72 views are uniformly generated after the model is aligned to the reference X-rays and the 3D volume is reconstructed using the FBP scheme. Due to the introduction of pose information, comparing the baseline models GRAF and MEDnerf, our proposed scheme has advantages. Meanwhile, the DRB module also improves the performance of 3D data reconstruction, but this also increases the inference time to some

Journal Pre-proof

extent. The time taken by the different models from fine-tuning to inference about the 0° to 360° view with their performance is shown in Fig.8. Compared with GRAF, the W/O DRB version of PGgraf adds the computation for viewpoint consistency judgment, which results in a slight increase in inference time. In addition, the parallel processing of different flipped features theoretically does not affect the computation time. PGgraf offers higher performance for all approximate time ranges. Depending on the practical considerations, different versions can be chosen.

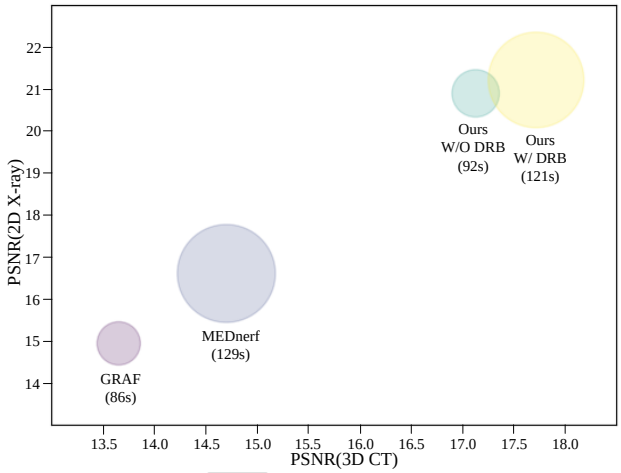


Figure 8: Runtime and performance of different models. Where the horizontal axis represents the quality of the 3D CT reconstructed using the FBP algorithm, the vertical axis represents the quality of the reconstructed X-rays of the new view, and the area represents the running time.

4.3.3. Quantitative Overhead Analysis of DRB

To clarify the computational cost introduced by the proposed Density Reconstruction Block (DRB), we quantitatively analyze its parameters, floating-point operations (FLOPs), and runtime overhead using standard tools. All calculations are consistent with the experimental setup in Section 4.2 (input resolution: 128×128, ray sampling count per pixel: 128, batch size: 16) to ensure fairness. The results are summarized in Table 9.

As shown in Table 9, the DRB module introduces minimal computational overhead to the baseline PGgraf framework. Specifically, it only adds 750 parameters (0.75K), accounting for merely 9.2% of the total model parameters,

Journal Pre-proof

Table 9: Quantitative overhead of the DRB module

Model Component	Parameters (K)	FLOPs (G)	Percentage of Model (Params/FLOPs)
PGgraf (w/o DRB)	8.19	14.32	-
PGgraf (w/ DRB)	8.94	15.76	-
DRB Module Alone	0.75	1.44	9.2% / 10.1%

and 1.44G FLOPs (10.1% of the baseline's total FLOPs). This lightweight design benefits from the efficient integration of multi-head attention and compact MLP layers, focusing solely on critical particle feature interaction without redundant computations. The marginal overhead is well-justified by significant performance gains: compared to PGgraf without DRB, the DRB-equipped model achieves an average PSNR improvement of 0.26–0.36 dB, a LPIPS reduction of 0.003–0.008 across all datasets and angle ranges, and more consistent 3D structure reconstruction in perpendicular views (Fig. 6). Consistent with Fig. 8, the DRB increases 0°–360° novel-view synthesis runtime by 31.5% (from 92s to 121s), an acceptable trade-off for clinical auxiliary diagnosis where enhanced reconstruction accuracy directly supports reliable decision-making. Compared to existing NeRF-based methods (e.g., MEDnerf and GRAF) plagued by structural distortion or unrealistic views, PGgraf achieves a favorable balance between computational efficiency and reconstruction quality via the lightweight DRB design.

5. Conclusion

The Neural Radiance Field (NERF) technique has achieved great success with natural images, and many studies are beginning to apply it to medical data reconstruction. A practical challenge is to generate successive novel-views based on a single X-ray. Compared to natural images, medical images are taken at a largely fixed angle with a standardized shooting pose. At the same time, under a fixed pose, similar medical images have an obvious structural prior. Based on the above considerations, we propose the Pose-Guided generative radiance field, named PGgraf. PGgraf combines ray direction and camera angle information with X-ray features, allowing the radiance field to reason about angles consistent with reality. Meanwhile, we designed a Density Reconstruction Block (DRB). It can combine all the particle density features sampled in a ray to predict the particle density in an integrated

Journal Pre-proof

way. In the fine-tuning stage, DRB can efficiently align the reference X-rays. PGgraf compares favorably with state-of-the-art NERF methods and offers significant advantages in imaging quality and visual perception. This work has great potential to aid physicians in diagnosis, reduce radiation damage to patients, and accelerate CT reconstruction.

PGgraf authorship contribution statement

Hangyu Li: Methodology, Software, Conceptualization, Writing-original draft. MoQuan Liu: Project administration, Supervision, Writing-review & editing. Nan Wang: Formal analysis, Data curation, Validation. Mengcheng Sun: Project administration, Supervision. Yu Zhu: Formal analysis, Supervision.

Conflict of interest

The authors declared no conflicts of interest with respect to the research, authorship, and publication of this paper.

Data availability statements

The data that support the findings of this study are available from the corresponding author, Yu Zhu, upon reasonable request.

Ethics statement

This work does not involve experimental procedures on human subjects or animals.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62476088 and 82502467 , and the Science and Technology Commission of Shanghai Municipality under Grant 20DZ22544000.

Journal Pre-proof

Journal Pre-proof

References

[1] P. Suetens, Fundamentals of medical imaging: Visualization for diagnosis and therapy, 2009. URL: https://api.semanticscholar.org/CorpusID:57401532.

[2] P. Lo, B. van Ginneken, J. M. Reinhardt, T. Yavarna, P. A. de Jong, B. Irving, C. I. Fetita, M. Ortner, R. Pinho, J. Sijbers, M. Feuerstein, A. Fabijańska, C. Bauer, R. R. Beichel, C. S. Mendoza, R. Wiemker, J. Lee, A. P. Reeves, S. Born, O. Weinheimer, E. M. van Rikxoort, J. Tschirren, K. Mori, B. Odry, D. P. Naidich, I. Hartmann, E. A. Hoffman, M. Prokop, J. J. H. Pedersen, M. de Bruijne, Extraction of airways from ct (exact'09), IEEE Transactions on Medical Imaging 31 (2012) 2093–2107. URL: https://api.semanticscholar.org/CorpusID:215761541.

[3] Y. Kasten, D. Doktofsky, I. Kovler, End-to-end convolutional neural network for 3d reconstruction of knee bones from bi-planar x-ray images, in: Machine Learning for Medical Image Reconstruction: Third International Workshop, MLMIR 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 8, 2020, Proceedings 3, Springer, 2020, pp. 123–133.

[4] T. Huynh, Y. Gao, J. Kang, L. Wang, P. Zhang, J. Lian, D. Shen, Estimating ct image from mri data using structured random forest and auto-context model, IEEE transactions on medical imaging 35 (2015) 174–183.

[5] S. Xie, W. Huang, T. Yang, D. Wu, H. Liu, Compressed sensing based image reconstruction with projection recovery for limited angle cone-beam ct imaging, in: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), IEEE, 2020, pp. 1307–1310.

[6] Y. Li, K. Li, C. Zhang, J. Montoya, G.-H. Chen, Learning to reconstruct computed tomography images directly from sinogram data under a variety of data acquisition conditions, IEEE transactions on medical imaging 38 (2019) 2469–2481.

Journal Pre-proof

Journal Pre-proof

[7] D. B. Lindell, J. N. Martel, G. Wetzstein, Autoint: Automatic integration for fast neural volume rendering, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 14556–14565.

[8] Y. Sun, J. Liu, M. Xie, B. Wohlberg, U. S. Kamilov, Coil: Coordinate-based internal learning for tomographic imaging, IEEE Transactions on Computational Imaging 7 (2021) 1400–1412.

[9] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, R. Ng, Nerf: Representing scenes as neural radiance fields for view synthesis, Communications of the ACM 65 (2021) 99–106.

[10] R. Martin-Brualla, N. Radwan, M. S. Sajjadi, J. T. Barron, A. Dosovitskiy, D. Duckworth, Nerf in the wild: Neural radiance fields for unconstrained photo collections, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 7210–7219.

[11] A. Yu, V. Ye, M. Tancik, A. Kanazawa, pixelnerf: Neural radiance fields from one or few images, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 4578–4587.

[12] Q. Wang, Z. Wang, K. Genova, P. P. Srinivasan, H. Zhou, J. T. Barron, R. Martin-Brualla, N. Snavely, T. Funkhouser, Ibrnet: Learning multi-view image-based rendering, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 4690–4699.

[13] A. Chen, Z. Xu, F. Zhao, X. Zhang, F. Xiang, J. Yu, H. Su, Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 14124–14133.

[14] M. Niemeyer, J. T. Barron, B. Mildenhall, M. S. Sajjadi, A. Geiger, N. Radwan, Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5480–5490.

Journal Pre-proof

Journal Pre-proof

[15] J. Kulhánek, E. Derner, T. Sattler, R. Babuška, Viewformer: Nerf-free neural rendering from few images using transformers, in: European Conference on Computer Vision, Springer, 2022, pp. 198–216.

[16] J. Yang, M. Pavone, Y. Wang, Freenerf: Improving few-shot neural rendering with free frequency regularization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 8254–8263.

[17] A. Trevithick, B. Yang, Grf: Learning a general radiance field for 3d scene representation and rendering (2020).

[18] K. Schwarz, Y. Liao, M. Niemeyer, A. Geiger, Graf: Generative radiance fields for 3d-aware image synthesis, Advances in Neural Information Processing Systems 33 (2020) 20154–20166.

[19] A. Corona-Figueroa, J. Frawley, S. Bond-Taylor, S. Bethapudi, H. P. Shum, C. G. Willcocks, 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), IEEE, 2022, pp. 3843–3848.

[20] J. Wynn, D. Turmukhambetov, Diffusionerf: Regularizing neural radiance fields with denoising diffusion models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 4180–4189.

[21] M. Qin, W. Li, J. Zhou, H. Wang, H. Pfister, Langsplat: 3d language gaussian splatting, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 20051–20060.

[22] Z. Chen, L. Yang, J.-H. Lai, X. Xie, Cunerf: Cube-based neural radiance field for zero-shot medical image arbitrary-scale super resolution, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 21185–21195.

[23] R. Zha, Y. Zhang, H. Li, Naf: neural attenuation fields for sparse-view cbct reconstruction, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2022, pp. 442–452.

Journal Pre-proof

[24] Y. Fang, L. Mei, C. Li, Y. Liu, W. Wang, Z. Cui, D. Shen, Snaf: Sparse-view cbct reconstruction with neural attenuation fields, arXiv preprint arXiv:2211.17048 (2022).

[25] N. Max, Optical models for direct volume rendering, IEEE Transactions on Visualization and Computer Graphics 1 (1995) 99–108.

[26] L. Mescheder, A. Geiger, S. Nowozin, Which training methods for gans do actually converge?, in: International conference on machine learning, PMLR, 2018, pp. 3481–3490.

[27] X. Pan, X. Zhan, B. Dai, D. Lin, C. C. Loy, P. Luo, Exploiting deep generative prior for versatile image restoration and manipulation, IEEE Transactions on Pattern Analysis and Machine Intelligence 44 (2021) 7474–7489.

[28] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, S. Paul Smolley, Least squares generative adversarial networks, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2794–2802.

[29] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).

[30] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, et al., The cancer imaging archive (tcia): maintaining and operating a public information repository, Journal of digital imaging 26 (2013) 1045–1057.

[31] A. A. Ali, S. S. Shalhoub, A. J. Cyr, C. K. Fitzpatrick, L. P. Maletsky, P. J. Rullkoetter, K. B. Shelburne, Validation of predicted patellofemoral mechanics in a finite element model of the healthy and cruciate-deficient knee, Journal of biomechanics 49 (2016) 302–309.

[32] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 586–595.

Journal Pre-proof

Highlights

PGgraf: Pose-Guided generative radiance field for novel-views on X-ray

•Synthesizing novel-views on X-rays that are consistent with realistic perspectives: We use the relative angle(pose) to the human body when taking an X-ray as the supervisory information of the radiance field, and propose the Pose-Guided generative radiance field (PGgraf), which can synthesize a continuous novel-views consistent with the realistic perspective by referring to a single X-ray.

•Pose information fusion strategies at different levels: We designed two different pose information fusion schemes; (1) The set of ray directions corresponding to each pixel in a patch as pose information. (2) The pose of the camera when taking an X-ray. The two scenarios have different information densities, both of which provide gains to the model.

•Density Reconstruction Block (DRB): Considering that samples of the same pose from different patients have similarities, we designed a Density Reconstruction Block (DRB) in the radiance field. DRB can make a comprehensive determination of particle density based on particle characteristics at different locations. The method allows efficient model fine-tuning in reference to a single X-ray using a priori information from the same site.

•Our proposed method, PGgraf, outperforms state-of-the-art methods in visual perception and image quality, as demonstrated by qualitative-quantitative experiments on both public and private datasets. Across the three angle ranges of 0° to 360°, -15° to 15°, and 75° to 105°, PSNR improved by an average of 4.18Db, SSIM improved by an average of 0.074, and LPIPS improved by an average of 50.7%.

Journal Pre-proof

Journal Pre-proof

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Journal Pre-proof