Non-local Guided Neural Fields for 4D CT Reconstruction

Qingyang Zhou, Yunfan Ye*, Zhihuang Liu, Chang Liu, and Zhiping Cai, *Member, IEEE*

***Abstract*—Dynamic CT reconstruction plays a crucial role in both medical and industrial applications. However, existing 4D CT reconstruction methods typically rely on complex regularization techniques or external large-scale training datasets, posing challenges for reconstruction quality and generalization when handling complex object motion and varied imaging modes. Neural Radiance Fields (NeRF) offer a promising approach to dynamic CT reconstruction, but existing NeRF-based methods often assume that the scene is low-rank, limiting their representation capabilities. To address these issues, we propose NG-NeRF. First, we combine 3D and 4D hash grids for scene representation, effectively reducing temporal redundancy in static regions of dynamic scenes while improving the model's representation capabilities and efficiency. Next, we design a non-local hash attention module to establish non-local dependencies between the features of different hash grids. This guides the model to adaptively select features based on hash table load information, significantly alleviating hash collisions and achieving the decoupling of dynamic and static regions. Besides, we introduce global continuity by employing mask positional encoding, which helps reduce the noise often introduced by grid features. Our experimental results on medical and industrial datasets demonstrate that the proposed method outperforms existing state-of-the-art methods by 5.84 dB and 3.4 dB, respectively, and exhibits excellent generalization ability across different 4D CT scenarios.**

***Index Terms*—4D CT reconstruction, neural radiance fields, hash grid.**

I. Introduction

COMPUTED Tomography (CT) is widely known for its ability to reveal the internal structure of objects using X-ray penetration [1]–[3]. 4D CT reconstruction can further provide the motion process of an object by introducing the time dimension. This technology holds significant potential for applications in medical radiotherapy [4], industrial product testing, and material analysis [5].

Nonetheless, 4D CT reconstruction is a challenging and inherently ill-posed problem. In medical scenarios, object motion is primarily caused by human breathing. During the acquisition of projections, motion state (phase) signals are typically captured using motion sensors. Vanilla 4D CT [6] reconstruction classifies projections into distinct phases according to motion signals and reconstructs the CT images for each phase individually. However, due to the small number of projections in different phases, streak artifacts are prone to occur [7].

Traditional 4D CT reconstruction methods employ regularization techniques [8] and deformation vector fields (DVFs) [9]–[11] to mitigate streak artifacts. However, excessive regularization and the complexity of hyperparameter adjustments hinder the diversity of motion representation [12]. Learning-based methods leverage large artifact-free datasets to train artifact-removal neural networks in the image domain [13]–[15], further enhancing reconstruction performance. Yet, these methods lack geometric understanding and have limited generalization capabilities. Furthermore, both methods above depend on sensors to capture motion states and are primarily suited for periodic motion in medical scenarios. In industrial scenarios, projection acquisition can only be obtained in a time-sequential manner, and it is impractical to use sensors to capture motion signals, so these methods are difficult to apply to industrial scenarios [5]. Hence, there is a critical need for a universal method capable of achieving high-quality dynamic CT reconstruction across diverse scenarios.

Recently, Neural Radiance Fields (NeRF) [16] has shown remarkable performance in the task of Novel View Synthesis (NVS) in natural scenes. This technology is highly flexible and scalable, and does not rely on additional datasets, providing a new solution for CT reconstruction. However, due to the impenetrability of natural light, reconstruction tasks in natural scenes primarily focus on the surface information of objects, so the design of existing NeRF methods is usually based on the low-rank assumption. In contrast, CT scenes require more dense spatial information, imposing higher demands on the model's representational capacity (as shown in Figure 1). Under these conditions, the low-rank assumptions adopted by many existing NeRF-based approaches significantly limit the model's ability to capture fine-grained spatial structures in high-resolution CT reconstructions. As a result, current NeRF-based 4D CT reconstruction methods [5], [17]–[19] are prone produce loss of volumetric details and are typically constrained to low-resolution reconstruction.

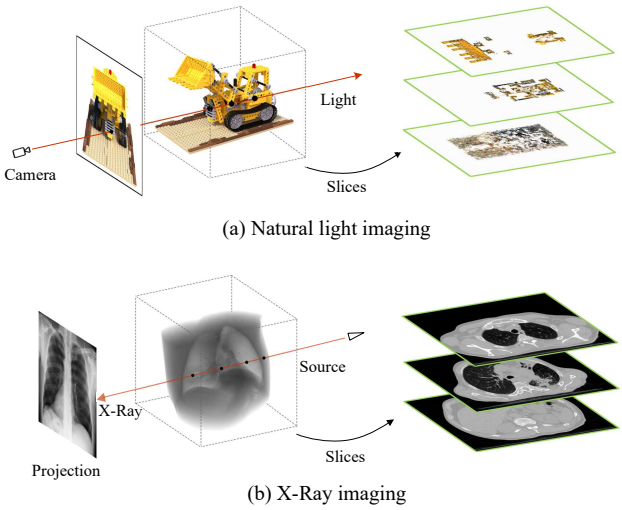
To address the above problems, we propose a novel framework, NG-NeRF. **First**, we adopt a combination of 3D and 4D hash grids [20] to represent 4D CT scenes, with the expectation that the 3D hash grid captures the static regions and prevents information redundancy. However, previous studies [21] show that a simple combination easily leads to the 4D hash grid overtaking the static region. In addition, hash collisions can

This work is supported by the National Natural Science Foundation of China (62402171, 62402505, 62472434), the National Key Research and Development Program of China (2022YFF1203001), the Sci-Tech Innovation 2030 Agenda(2023ZD0508600), and the Science and Technology Innovation Program of Hunan Province (2022RC3061).

Q. Zhou, Z. Liu, C. Liu and Z. Cai are with the College of Computer Science and Technology, National University of Defense Technology, Changsha 410073, China (e-mail: z_qy@nudt.edu.cn; lzhliu@nudt.edu.cn; liudawn@nudt.edu.cn; zpcai@nudt.edu.cn).

Y. Ye is with the College of Design, Hunan University, Changsha 410082, China (e-mail:yeyunfan@hnu.edu.cn).

*Corresponding author is Yunfan Ye.



(a) Natural light imaging

(b) X-Ray imaging

Fig. 1. Comparison between natural light imaging and X-ray imaging. (1) Natural light imaging, constrained by the impenetrability of natural light, captures only the surface details of an object. As shown in the slices, internal information remains unclear and ambiguous. (b) In contrast, CT imaging focuses on capturing detailed information across the entire imaging space.

further degrade reconstruction quality. We argue that both information redundancy and hash collisions fundamentally arise from an imbalance in hash table load. Therefore, **secondly**, we propose a non-local hash attention module that establishes closer non-local dependencies between the features of different hash grids corresponding to the sampling points. This allows the model to adaptively balance the load of the hash table, alleviating hash collisions, improving feature utilization, and making the model more compact, while decoupling dynamic and static regions. **Besides**, we introduce Mask Positional Encoding to guide the model in learning global information, implicitly smoothing images without relying on regularization terms, and further mitigating artifacts caused by hash collisions and grid features.

We conduct experiments on high-resolution medical and industrial datasets with different scenarios and scanning modes. Experimental results demonstrate that our method consistently outperforms current mainstream approaches across all datasets. The main contributions of this work are summarized as follows:

- We propose NG-NeRF, a novel framework capable of achieving high-quality 4D CT reconstruction across various scenarios and scanning modes without requiring additional datasets.
- We introduce a combination of 3D and 4D hash grids to represent 4D CT scenes, design a non-local hash attention module and incorporate mask positional encoding. These approaches enhance the model's representational capacity, significantly mitigate hash collisions using non-local information, and maintain the model's compactness and efficiency.
- Experiments on challenging high-resolution 4D CT datasets demonstrate that NG-NeRF significantly outperforms existing methods and has high practical value.

The subsequent sections of this paper are organized as follows. In § II, we review related works about 4D CT reconstruction, neural fields for natural scenes, and neural fields for CT reconstruction. In § III, we introduce the preliminaries of NeRF-based 4D CT reconstruction and describe the proposed method in detail. In § IV, we present our experimental results and analysis. In § V, we conclude this paper in the end.

II. Related Work

A. 4D CT reconstruction

In traditional iterative 4D CT reconstruction algorithms, motion is typically represented as deformation vector fields (DVFs) to compensate for an object's motion trajectory during scanning. DVFs are generally derived through two approaches: the first is based on non-rigid registration between 4D CT images across different phases [9]–[11]; the second directly calculates DVFs from projections [22], [23]. However, registration often introduces errors, and optimizing DVFs remains challenging. To address this issue, some studies [24], [25] have proposed incorporating regularization terms, which leverage prior information to improve DVF estimation. Nevertheless, excessive reliance on regularization can hinder the algorithm's ability to handle diverse motion patterns and multiple scanning modes effectively [12].

Learning-based reconstruction methods often train a post-processing model [13]–[15] to enhance the quality of reconstructed images by removing artifacts and noise or improving detail resolution. These methods has the advantage of fast speed, but due to the limitation of computing power, the receptive field of the neural network is difficult to cover 4D images. On the other hand, [26], [27] directly learn a mapping from the projections to the 4D reconstructed images. Despite their potential, learning-based methods heavily rely on dataset diversity and lack generalization to unknown data, thereby limiting their practical applicability.

We argue that the community requires a method capable of achieving high-quality reconstruction across various scenarios. The approaches mentioned above face challenges in meeting the dual demands of reconstruction accuracy and generalization.

B. Neural Fields for Natural Scene

NeRF [16] was the first to combine volume rendering with neural networks, using a multi-layer perceptron (MLP) to represent 3D scenes. Many subsequent works have built upon this foundation, making significant improvements and applying them to various downstream tasks [28]–[30]. For example, Müller et al. [20] proposed representing the scene as multiresolution grid features and mapping them into a hash table, significantly enhancing training and rendering speeds. Zhang et al. [31] incorporated hash grids to address the limitations of existing methods in terms of pose and radiation field consistency. Additionally, several other improvements based on NeRF have emerged, including sparse view [32]–[34], accelerated rendering [35], [36], and reduced memory usage reducing [37], [38].

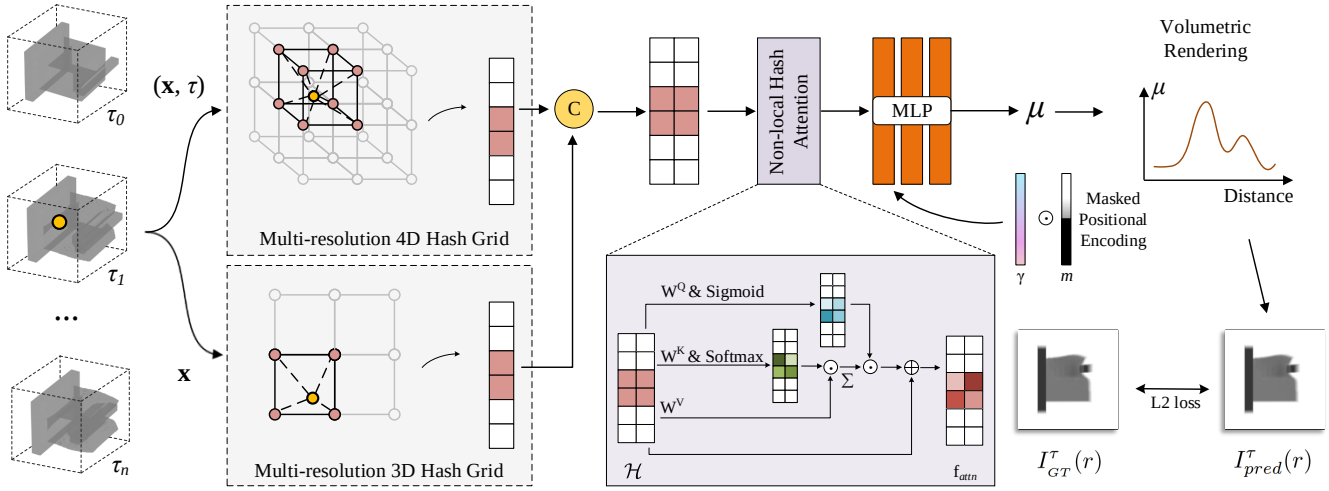


Fig. 2. Overview of our method. First, we input the 4D coordinates $(\mathbf{x}, \tau)$ and 3D coordinates $\mathbf{x}$ of the sampling points into 3D and 4D hash grids (Dimensionality reduction on 3D and 4D hash grids are performed in the figure for the convenience of representation), respectively, and use hash indexing and linear interpolation to extract the grid features. Then, these grid features are concatenated and passed through the Non-local Hash Attention module to establish non-local dependencies. Next, we reshape the output features into 1D and concatenate them with the mask positional encoding, and then input them into the tiny MLP to obtain the attenuation coefficient μ. We predict μ of the sampling point on the ray and obtain the predicted projection $I^{\tau}_{pred}(r)$ using volumetric rendering as defined in Eq. 3. Finally, the $I^{\tau}_{pred}(r)$ is compared with the actual measured projection $I^{\tau}_{GT}(r)$, and the model is optimized using L2 loss.

To accommodate the ever-changing nature of the world, many studies have extended NeRF to dynamic scenes [39], [40]. Existing dynamic NeRF methods can be divided into two categories. The first category is based on the concept of decoupling, which represents the dynamic scene as a canonical scene and a deformed scene [41]–[44], or directly decouples the dynamic scene into static and dynamic regions [21], [45], and are represented using independent models respectively. These methods are typically implemented using MLPs, which result in low computational efficiency. Among them, [45] introduced hash encoders to reduce training time, but its results showed that it was less effective than the MLP-based implementation. The second category of methods decomposes 4D scenes into low-dimensional planes [46], [47], enhancing scene representation compactness and efficiency.

We believe that incorporating explicit grid features would enhance the representation of dense scenes. Existing methods encode explicit features through tensor decomposition [45], [46] or hash mapping [21] to compactly represent 4D scenes and improve computational efficiency. However, in 4D CT scenarios, the utilization of features still needs to be further optimized to improve reconstruction performance.

C. Neural Fields for CT Reconstruction

Recently, several studies have applied NeRF to CT reconstruction [48]–[52], demonstrating the great potential of this method in the field. However, NeRF still faces certain challenges in 4D CT scenes. Leveraging the decoupling concept in dynamic NeRF, [5], [18], [19] have modeled dynamic CT scenes as a canonical scene and a time-varying deformation scene. However, since the canonical scene and deformation scene are jointly optimized, both of them may be distorted simultaneously, leading to suboptimal solutions. Therefore, these methods usually requires the introduction of prior knowledge or regularization to constrain the model. Maas et al. [17] follows the idea of [21] to decouple the dynamic and static regions of the 4D scene into separate radiation fields, and introduces Factorization loss to ensure the correct separation. This method is similar to our idea. However, it remains questionable whether the strict separation of dynamic and static components contributes to improving image quality. In addition, most of the existing NeRF-based methods are implemented based on MLPs, which may have limited representation power, often neglecting detailed information, and are thus restricted to lower-resolution reconstructions.

III. Methodology

A. Preliminaries

In dynamic natural scenes, NeRF typically learns a mapping from 3D coordinates, viewing direction and time to color and density for NVS. Unlike natural scenes, only isotropic density (attenuation coefficient) is of interest in CT imaging, so there is no information about viewing direction and color. The imaging principle is illustrated in Figure 1-b. In CT imaging, an X-ray source emits a cone beam of rays. As the rays pass through the scanned object, their energy is attenuated. A flat-panel detector then captures the attenuated energy, forming a projection. Based on Beer law [53], the measured projection can be expressed as:

$$I^{\tau}_{GT}(r) = I_0 \exp\left(-\int_{h_n}^{h_f} \mu(r(h), \tau) dh\right), \tag{1}$$

where $I^{\tau}_{GT}(r)$ and I_0 represent the ray projection value detected at time τ and ray source intensity, respectively, h_n and h_f represent the near and far ends where the ray intersects the

reconstruction space. $r(h) = o + hd$ represents the coordinate on the ray. $\mu(r(h), \tau)$ represents the attenuation coefficient of the coordinate $r(h)$ at time τ. Our goal is to learn a function from the coordinate $\mathbf{x} \in \mathbb{R}^3$ and time $\tau \in \mathbb{R}$ to the attenuation coefficient $\mu \in \mathbb{R}$, so the model can be expressed as:

$$\mathcal{F}_\Phi : (\mathbf{x}, \tau) \rightarrow \mu, \tag{2}$$

where Φ represents learnable parameters. We discretize Eq. 1 and substitute it into Eq. 2 to derive the discrete volumetric rendering formula, through which the predicted projection value $I^\tau_{pred}(r)$ is obtained. The discrete volumetric rendering can be expressed as:

$$I^\tau_{pred}(r) = I_0 \exp(-\sum_{i=0}^{M} F_\Phi(r(h_i), \tau)\delta_i), \tag{3}$$

where $\delta_i = \|r(h_{i+1}) - r(h_i)\|_2$ represents the distance between sampling points i and $i+1$. Unlike natural scenes, the attenuation coefficients at various positions in the reconstruction space of CT scenes have similar importance. Therefore, we employ layered sampling to evenly distribute the sampling points along the ray. Finally, we use the L2 Loss between the predicted and the measured projection value as the loss function to train the model:

$$\mathcal{L} = \sum_{r \in \mathcal{B}} \left\| I^\tau_{GT}(r) - I^\tau_{pred}(r) \right\|^2_2. \tag{4}$$

We do not incorporate any prior knowledge or regularization terms to ensure the model's generalization performance. It is important to emphasize that the ultimate goal of CT reconstruction is to accurately predict the attenuation coefficient distribution in the space, rather than novel view synthesis. The overview of our method can be seen in Figure 2.

B. Combined Hash Grids

To accelerate training, Müller et al. [20] proposed a multi-resolution hash grid, which has demonstrated strong performance in novel view synthesis and reconstruction tasks for natural scenes. The hash grid represents the space as voxel grids of l levels, each voxel grid stores F-dimensional features. Then, each level of the grid is mapped to a hash table through a spatial hash function [54]. The multi-resolution design of hash grids allows for efficient learning of high-frequency details in CT scenes while keeping the model compact.

Specifically, given any dimension coordinate $\mathcal{X}$, each level extracts the grid features containing the coordinates, and uses linear interpolation to obtain the output features, and then concatenates the features of each level. Let $D = F \cdot l$, the hash grid feature can be expressed as:

$$h(\mathcal{X}) = \left[h_{1,1}(\mathcal{X}), \ldots, h_{F,1}(\mathcal{X}), \ldots, h_{F,l}(\mathcal{X}) \right], h(\mathcal{X}) \in \mathbb{R}^D. \tag{5}$$

In the dynamic CT reconstruction task, a straightforward approach is to directly use a 4D hash grid to represent the 4D CT scene. However, the static parts of the scene consume a significant portion of the 4D hash grid features, resulting in information redundancy. For example, in a 4D CT reconstruction task with $\mathcal{T}$ frames, using a 4D hash grid to encode static regions would result in each static point being mapped to $\mathcal{T}$ hash table entries with the same value, leading to significant feature redundancy. Inspired by the dynamic NeRF decoupling method [21], we employ both 3D and 4D hash encoders to represent the 4D scene, with the 3D hash grid responsible for capturing static regions and reducing redundant feature storage. Specifically, given normalized spatial coordinates $\mathbf{x}$ and time τ, project them into two hash grids, and concatenate the output features:

$$\mathcal{H}(\mathbf{x}, \tau) = [h_{3D}(\mathbf{x}), h_{4D}(\mathbf{x}, \tau)], \mathcal{H}(\mathbf{x}, \tau) \in \mathbb{R}^{D \times 2}. \tag{6}$$

We expect the 3D hash grids to handle the task of representing static regions, but this cannot be guaranteed through simple feature concatenation alone. From another perspective, while the combined hash grid offers an excellent solution space, additional guidance is still needed to help the model find the optimal solution. Furthermore, hash collisions present a significant issue. Müller et al. [20] believe that in sparse natural scenes, hash collisions can be mitigated by incorporating hash grid features of different levels, which has minimal impact on the reconstruction quality. However, in dense 4D CT scenes, hash collisions significantly degrade the quality of the reconstructed image (as shown in Figure 3-a).

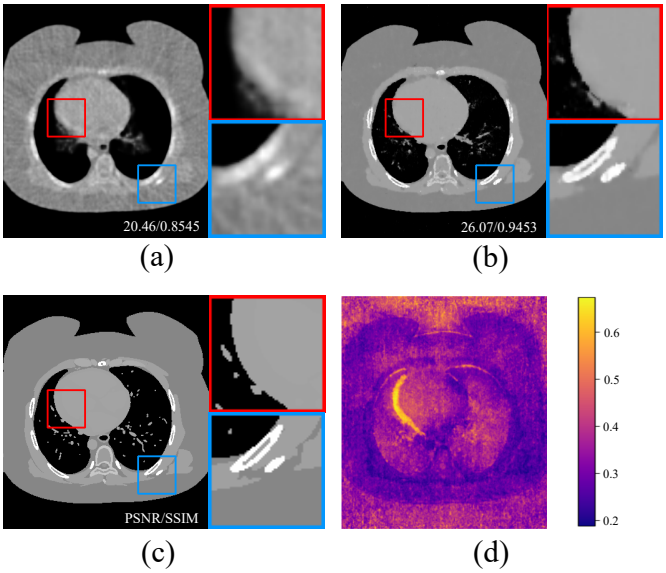


Fig. 3. Non-local Hash Attention (NHA) function analysis. (a) Reconstruction result without NHA. (b) Reconstruction result with NHA. (c) Ground truth. (d) Weight of 4D hash grid features.

C. Non-local Hash Attention

Using regularization techniques [17], [21] to strictly decouple the scene is a feasible solution. However, this requires complex hyperparameter adjustments, which hinder the generalization of the model and have limited effectiveness in alleviating hash collisions.

To address these issues, we propose Non-local Hash Attention (NHA), which leverages the non-local information from sample points in different hash grids to assist the model in adaptive feature selection. Unlike NLP or CV tasks, we only

need to apply self-attention between two hash grid features, and the sequence length is very short. However, due to the large number of sample points, minimizing computational complexity is crucial. We consider that the bottleneck of the efficiency of the short sequence self-attention is the dot product, so we aim to reduce the number of dot products as much as possible.

Inspired by AFT [55], first, let the combined hash grid feature of sample point $\mathbf{q} = (\mathbf{x}, \tau)$ be $\mathcal{H}(\mathbf{q}) \in \mathbb{R}^{D\times 2}$, T represents the number of hash grids ($T = 2$). After the features pass through three linear layers $\mathrm{W^Q}$, $\mathrm{W^K}$ and $\mathrm{W^V}$, we get the query, key and value:

$$\mathrm{Q} = \mathcal{H}(\mathbf{q})\mathrm{W^Q}, \mathrm{K} = \mathcal{H}(\mathbf{q})\mathrm{W^K}, \mathrm{V} = \mathcal{H}(\mathbf{q})\mathrm{W^V}, \tag{7}$$

where Q, K and V $\in \mathbb{R}^{D\times 2}$. Then, the Hadamard product is used to achieve efficient interaction between elements:

$$\mathrm{f}_{attn}(\mathbf{q}) = \sigma\left(\mathrm{Q}\right) \odot \sum_{t'=0}^{T} \left(Softmax(\mathrm{K}) \odot \mathrm{V}\right)_{t'} + \mathcal{H}(\mathbf{q}), \tag{8}$$

where $\odot$ represents the Hadamard product and $\sigma(\cdot)$ represents the $Sigmoid$ function. By applying $Softmax$ to the K and $Sigmoid$ to the Q, the contribution of the value at each position to the final output can be effectively adjusted.

Figure 3-d visualizes the weight of 4D hash grid features ($weight = Softmax(\mathrm{K})_{t'=1}$) to each sample point in the reconstructed image. It can be observed that the NHA tends to let the 4D hash grid take over the dynamic region (e.g., the left side of the heart and the top of the lungs, which have higher weights), while reducing the weight of the static region. This demonstrates its ability to automatically decouple dynamic and static regions. In addition, we also observed that NHA does not fully decouple the scene. We consider this because the NHA adaptively guides the model to optimize feature selection based on the load information of the hash table and the random hash collisions. This partial decoupling is only the result of optimizing feature selection, which, in contrast to strict decoupling [17], [21], provides greater flexibility in addressing artifacts caused by hash collisions.

Complexity Analysis. As can be seen from Eq.8, NHA completely eliminates the dot product, with its computational complexity stemming from two Hadamard products. Therefore, its complexity is $\mathcal{O}(TD)$. In comparison, multi-head attention [56] $\mathcal{O}(T^2D)$ and linear attention [57] $\mathcal{O}(TD^2)$ have higher complexity. In tasks with extremely short sequences and a large number of sample points, this approach can significantly reduce training time with almost minimal performance loss (see Table VIII).

D. Masked Positional Encoding

The feature units in the hash grids are usually optimized independently. Although the grid-based approach helps capture local details of the scene, it lacks global information and is prone to introduce noise, resulting in a locally optimal solution [58]. The vanilla NeRF encodes global 3D coordinates and inputs them into the same MLP for scene learning. This MLP captures global scene information and introduces continuity, but its representation ability in detail is insufficient.

To combine both coarse global and fine information, we introduce the positional encoding to encode the 4D coordinates, concatenate the result with $\mathrm{f}_{attn}(\mathbf{q})$, and input them into MLP (see Figure 2). Frequency encoding $\gamma(\cdot)$ can be expressed as:

$$\gamma_L(\mathbf{q}) = [\mathbf{q}, \cdots, \sin\left(2^{L-1}\pi\mathbf{q}\right), \cos\left(2^{L-1}\pi\mathbf{q}\right)], \tag{9}$$

where L is the encoding length. However, high-frequency positional encoding can also easily lead to the model being overfitted to noise. Our primary goal is to make the MLP learn low-frequency global information. Inspired by [32], [33], we introduced a mask that shields high-frequency inputs at the beginning of the iteration, guiding the model to focus on low-frequency information and gradually introducing high-frequency signals as the number of iterations increases. Let n and N represent the current iteration number and the total number of iterations, respectively. The mask frequency encoding can be expressed as:

$$\begin{aligned}\gamma'_L(n, N, \mathbf{q}) = [m_0(n)\mathbf{q}, ..., m_{L-1}(n)\sin(2^{L-1}\pi\mathbf{q}),\\ m_{L-1}(n)\cos(2^{L-1}\pi\mathbf{q})],\end{aligned} \tag{10}$$

$$with \quad m_i(n) = \begin{cases} 1 & \text{if} \quad i \le \frac{nL}{N} \\ \frac{nL}{N} - \lfloor \frac{nL}{N} \rfloor & \text{if} \quad \frac{nL}{N} < i \le \frac{nL}{N} + 1 \\ 0 & \text{if} \quad i > \frac{nL}{N} + 1 \end{cases} \tag{11}$$

The final predicted attenuation coefficient can be expressed as:

$$\mu(\mathbf{q}) = MLP([\mathrm{f}_{attn}(\mathbf{q}), \gamma'_L(n, N, \mathbf{q})]). \tag{12}$$

Mask positional encoding (MPE) introduces global information, implicitly smooths the image, helps alleviate noise and artifacts, and does not require complex regularization hyperparameter adjustments. In addition, MPE assists NHA with spatial localization, enabling it to adjust attention weights based on spatial context rather than relying solely on hash features.

IV. Experiments

A. Dataset Description

1) Medical Datasets: In medical scenarios, object movement primarily arises from periodic breathing. During dynamic scanning, the motion state of the projections can be determined using a sensor and assigned to different phases, as illustrated in Figure 4-a. We collected chest high-resolution 4D CT datasets, including 4D extended cardiac torso (XCAT) [59] phantom and 4D-Lung Cancer Imaging Archive (TCIA) [60], which are consistent with [13]. Each 4D CT case features a coronal resolution of 512×512, an axial resolution ranging from 70 to 160, and a temporal resolution of 10 (corresponding to 10 respiratory phases). We simulated the scanning mode of the Varian Medical System and synthesized 100 projections within a 360° scanning range, where each phase contains 10 projections at different angles.

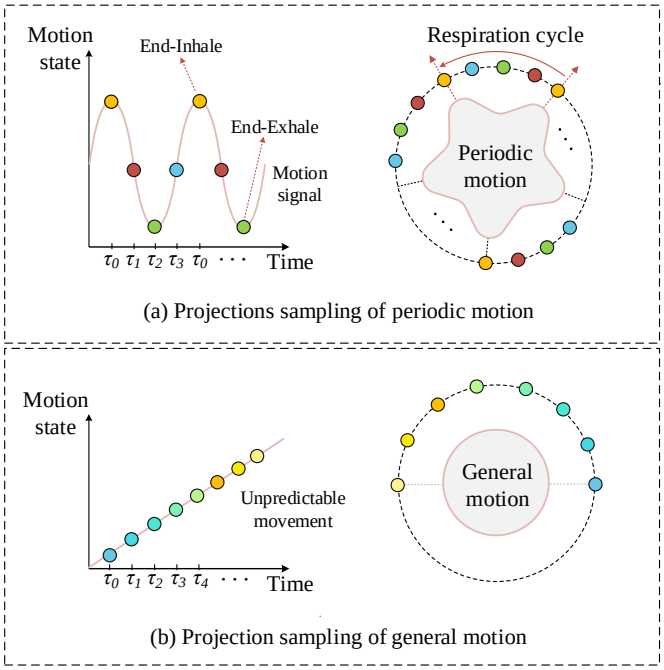


Fig. 4. Schematic diagram of scanning modes and projection sampling. (a) In the medical dataset, due to the periodicity of motion and phase gating technology, the same state can be sampled from multiple angles. (b) In the industrial dataset, as the object's motion state cannot be measured, sampling is performed sequentially over time.

TABLE I
HYPERPARAMETERS OF HASH GRIDS.

Hyperparameters	3D	4D
Number of levels	16	16
Hash table size	2^{19}	2^{20}
Feature dimension	2	2
Base resolution	16^3	$16^3 \times 15$
Finest resolution	2049^3	$2049^3 \times 300$

2) Industrial Datasets: To verify the generalization ability of our method, we further investigated its reconstruction performance on general object motion. To this end, we collected 4D CT datasets[1] of damage and evolution over time of aluminum products under external forces [5], which holds significant value for analyzing the performance of materials. Each set of CT images in the dataset features a spatial resolution of 256^3 and a temporal resolution of 60 (corresponding to 60 different motion states). We synthesized 240 projections within a 180° scanning range, where each motion state contains 4 projections with limited angles. Notably, due to the unpredictability of motion, the projections were labeled in the order of acquisition time, as illustrated in Figure 4-b. Compared to the medical dataset, the CT images in industrial dataset contain relatively less high-frequency spatial information but exhibit higher temporal resolution. Additionally, the limited viewing angles pose greater challenges for reconstruction. We provide video demonstrations of various sampling modes in the Supplementary Material.

[1]https://library.ucsd.edu/dc/object/bb74156780

B. Implementation Details

Our method is implemented using PyTorch and CUDA frameworks. The number of iterations per scene is set to 40k, and each iteration processes 1024 rays. Depending on the resolution of the reconstructed image, the number of sampling points per ray is set to 768 for the medical datasets and 320 for the industrial datasets. We use the Adam optimizer, set the initial learning rate to 1e-2, and gradually reduce it to 1e-6 through cosine decay. At the same time, a warm-up mechanism is added to improve the stability in the early stage of training. All experiments are completed on a single RTX 4090 GPU.

The hyperparameter settings for the hash grids are detailed in Table I. In the Mask positional encoding, L is set to 6. The tiny MLP consists of 4 fully connected layers with 32 channels, and each layer uses the Softplus activation function. We use Peak Signalto-Noise Ratio (PSNR) and Structural Similarity (SSIM) to evaluate the quality of the reconstructed images.

C. Results and Comparison

1) Results in Medical Datasets: In the medical dataset, we compare NG-NeRF with various state-of-the-art 4D CT reconstruction methods, including traditional methods (Gate-FDK [6], PICCS [25]), learning-based methods (CycN-Net [13], TT U-Net [61]), and NeRF-based methods (K-plane [47], Hex-plane [46], INR [5], and NeRF-CA [17]). For a fair of comparison, the training set of TT U-Net is consistent with CycN-Net. Additionally, K-plane and Hex-plane were originally designed for NVS in natural scenes. For CT reconstruction, we adapted them by removing viewing direction and color-related components while retaining the regularization techniques.

The quantitative reconstruction results are presented in Table II. As shown, our method significantly outperforms existing methods. Compared to the best NeRF-based method (NeRF-CA), our method improves PSNR and SSIM by 5.84 dB and 9.71%, respectively. Compared with the best learning-based method (TT U-Net), our method improves PSNR and SSIM by 11.31 dB and 10.94%, respectively.

We visualize the reconstruction results of XCAT and two cases in TCIA at the End-Inhale and End-Exhale in Figure 5. It can be observed that, in the traditional method, PICCS employed regularization technology to reduce streak artifacts, but the resulting image remains relatively blurry.

In the learning-based method, TT U-Net can reproduce clear image details, but the images are darker overall. This is because these methods do not reconstruct based on the principles of CT imaging, leading to inaccurate HU predictions, which is crucial information in medical diagnosis. In fact, the generalization problem of learning-based methods is not only reflected in the difficulty of applying them to unknown datasets. Different numbers of perspectives and different reconstruction preprocessing (Gate-FDK) implementation frameworks also have a significant impact on the results. We tried our best to eliminate the impact of generalization, but there are still problems with inaccurate HU value prediction and incomplete artifact removal.

TABLE II
QUANTITATIVE EVALUATION RESULTS ON MEDICAL DATASETS. THE BEST AND SECOND-BEST METHODS ARE HIGHLIGHTED BY **BOLD** AND UNDERLINE, RESPECTIVELY.

Method	XCAT		Case 1		Case 2		Case 3		Case 4		Average	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Gate-FDK	10.53	0.2609	12.79	0.2872	10.72	0.2622	7.250	0.2008	12.71	0.2542	10.86	0.2531
PICCS	21.72	0.5652	19.24	0.5089	21.74	0.5409	19.46	0.4995	19.12	0.5011	20.25	0.5231
CycN-Net	18.40	0.8046	19.66	0.8277	20.15	0.7670	20.34	0.7551	17.62	0.7343	19.23	0.7777
TT U-Net	20.26	0.8992	17.64	0.8159	18.86	0.8107	18.98	0.8172	17.55	0.7867	18.65	0.8259
Hex-plane	17.13	0.5458	22.40	0.6121	22.17	0.6234	23.46	0.6403	22.10	0.5854	21.45	0.6014
K-plane	21.49	0.6897	23.09	0.6367	24.01	0.6339	23.08	0.6913	21.40	0.6314	22.61	0.6566
INR	21.77	0.8111	24.67	0.7379	20.45	0.6926	19.24	0.6555	21.49	0.6361	21.52	0.7066
NeRF-CA	20.40	0.8508	25.66	0.8337	27.04	0.8565	23.45	0.8394	24.04	0.8106	24.12	0.8382
Ours	**26.07**	**0.9453**	**30.43**	**0.9290**	**31.52**	**0.9388**	**31.27**	**0.9380**	**30.49**	**0.9254**	**29.96**	**0.9353**

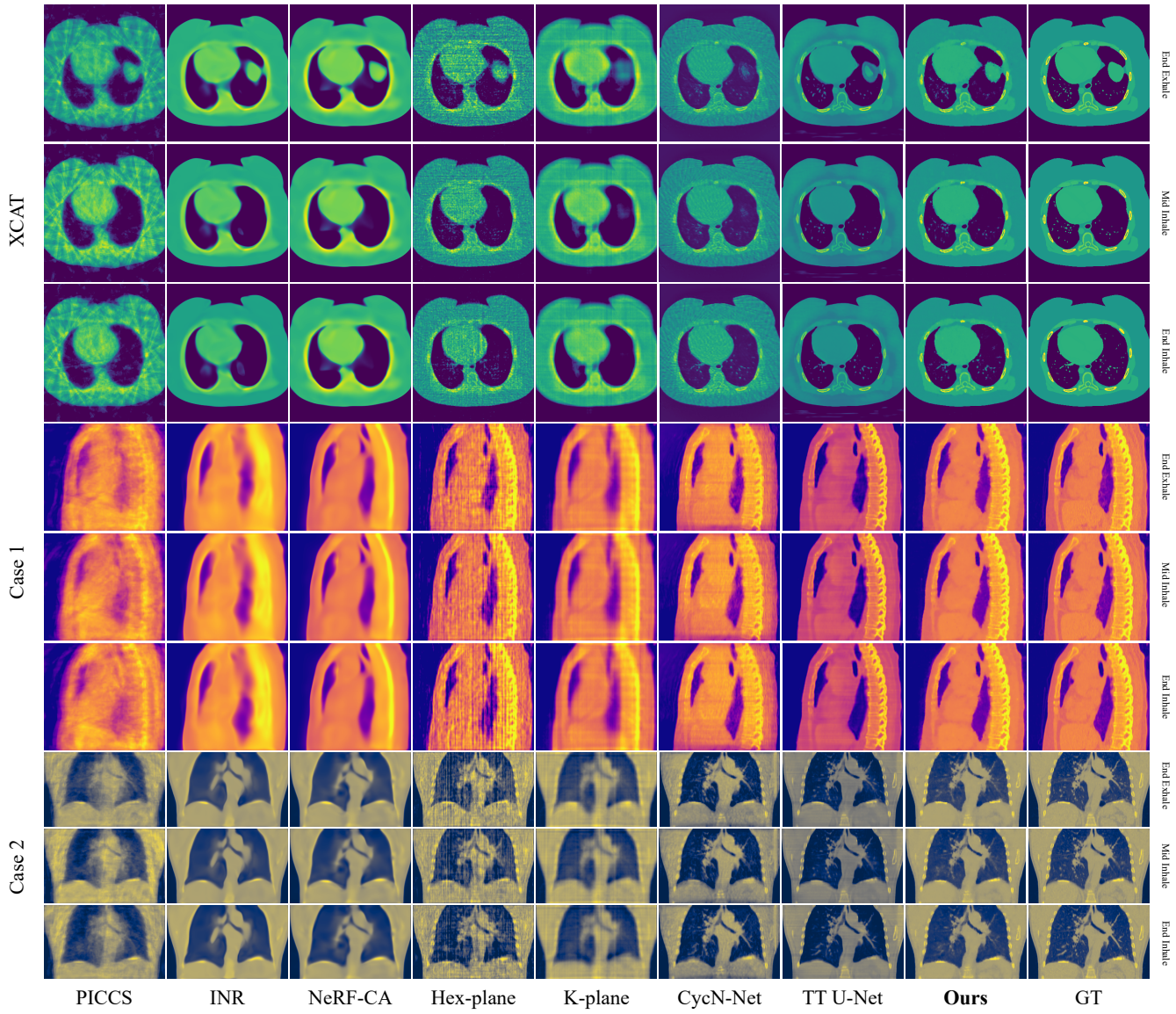


Fig. 5. Visualizing reconstruction results of medical datasets. We present the reconstruction results at the End-Exhale, Mid-Inhale and End-Inhale of the breathing cycle for different cases, with distinct colors used to differentiate between cases. The display window is set to a range of [-1000, 500] HU.

In the NeRF-based method, K-plane and Hex-plane represent 4D scenes based on grid features. Since the grid features are optimized independently, there is significant noise. INR and NeRF-CA represent 4D scenes using MLPs, which capture global information and produce smoother reconstructions, but they lack high-frequency details and require extensive hyperparameter tuning for optimal reconstruction quality. In contrast, our method outperforms existing methods in both

TABLE III
QUANTITATIVE EVALUATION RESULTS ON INDUSTRIAL DATASETS. THE BEST AND SECOND-BEST METHODS ARE HIGHLIGHTED BY **BOLD** AND UNDERLINE, RESPECTIVELY.

Method	Alum 1		Alum 2		Alum 3		Alum 4		Alum 5		Average	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Hex-plane	19.37	0.8087	19.63	0.8133	22.52	0.8295	21.38	0.8420	18.58	0.8181	20.29	0.8223
K-plane	20.01	0.9310	19.87	0.9389	21.93	0.9455	23.09	0.9656	18.21	0.9384	20.62	0.9439
INR	22.73	0.9379	19.71	0.9066	21.56	0.9100	23.12	0.9457	20.22	0.9155	21.47	0.9231
NeRF-CA	22.33	0.9630	21.27	0.9671	26.51	0.9766	23.99	0.9747	18.84	0.9462	22.59	0.9655
Ours	**26.49**	**0.9788**	**24.71**	**0.9746**	**28.85**	**0.9827**	**27.77**	**0.9821**	**22.15**	**0.9622**	**25.99**	**0.9761**

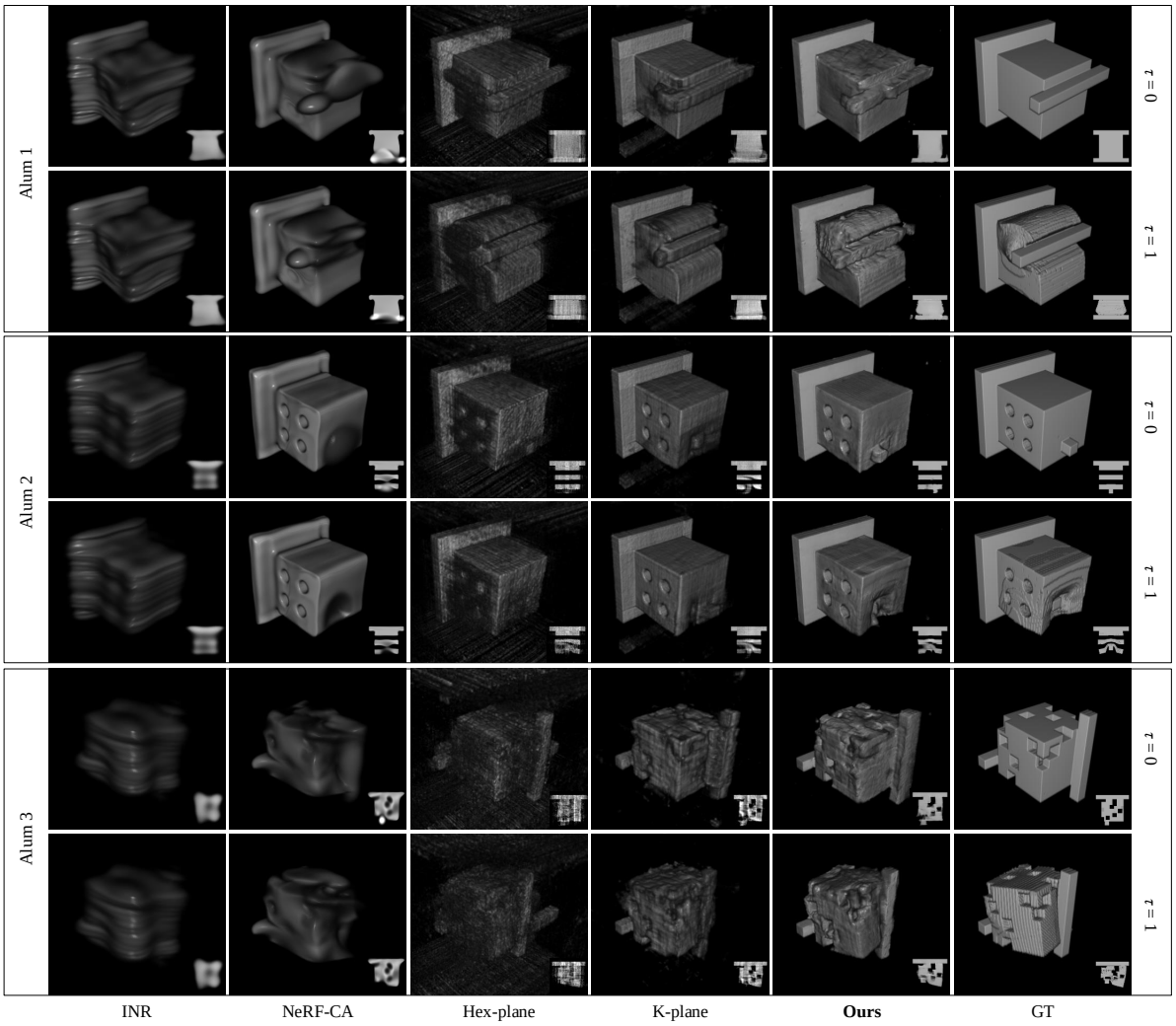


Fig. 6. Visualizing reconstruction results of industrial datasets. We present the rendering results at the beginning ($\tau = 0$) and end ($\tau = 1$) of the deformation process of aluminum products, accompanied by slice results in the bottom right corner.

local detail preservation and overall visual effects.

2) Results in Industrial Datasets: Since motion signals cannot be perceived, we excluded traditional and learning-based methods that rely on motion signals from the comparison. The quantitative results are shown in Table III, which can be seen that our method outperforms the second-best method (NeRF-CA) by 3.4 dB in PSNR and 1.06% in SSIM.

Compared to the medical dataset, the 4D CT images in the industrial dataset contain less high-frequency spatial information, and the implicit smoothing characteristics of the MLP-based methods (INR and NeRF-CA) give them certain advantages in this dataset. However, as can be seen from the visualization results in Figure 6, these methods still struggle to capture fine details at the edges. We present dynamic reconstruction results in the Supplementary Material.

3) Results in Natural Scenes: **4D Natural Scenes.** To investigate the potential of NG-NeRF in 4D natural scenes, we partially modified its model architecture and sampling strategy. Specifically, we introduced a viewing direction-dependent color output branch and adopted the coarse-to-fine sampling

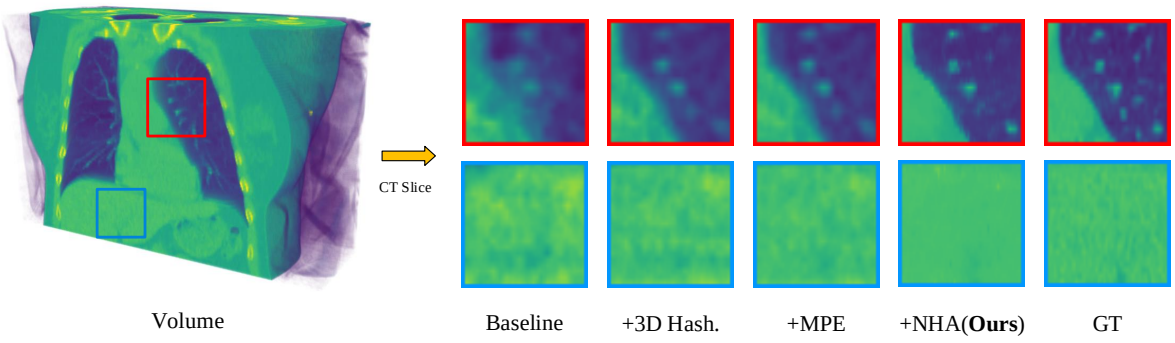


Fig. 7. Visual analysis of key components.

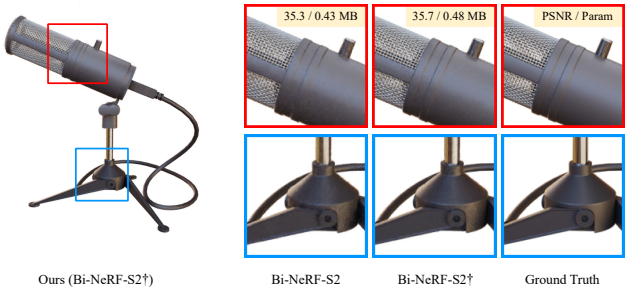


Fig. 8. Visualizing reconstruction results of Synthetic-NeRF datasets. The symbol †indicates that NHA is used. Incorporating NHA results in smoother rendered images while introducing only a small number of additional parameters.

TABLE IV
QUANTITATIVE EVALUATION RESULTS ON D-NERF DATASETS. THE BEST AND SECOND-BEST METHODS ARE HIGHLIGHTED BY **BOLD** AND UNDERLINE, RESPECTIVELY.

Method	PSNR	SSIM	LIPIS
Hex-plane	31.01	0.9679	0.0495
K-plane	**31.57**	**0.9713**	**0.0427**
Ours	31.06	0.9681	0.0452

scheme to enable it to be applied to natural scenes. The model was evaluated on the D-NeRF dataset [41], with both quantitative and qualitative results presented in Table IV.

The results indicate that our method does not outperform K-Plane and Hex-Plane. We attribute this to two main factors: First, in the novel view synthesis (NVS) task for sparse natural scenes, the effect of hash collisions is minimal. The multi-resolution design of the hash grid is already sufficient to mitigate collisions, which limits the effectiveness of our proposed NHA. Second, K-Plane and Hex-Plane incorporate low-rank regularization through tensor decomposition during optimization, constraining the parameter space to avoid overfitting and thereby achieving better synthesis quality [35]. In contrast, NG-NeRF is better suited for reconstructing high-rank scenes and shows limited advantage in the NVS task.

3D Natural Scenes. In the study of NeRF compression, similar combined hash grid strategies have been employed. For example, Bi-NeRF [37] utilizes 2D–3D hybrid binary hash grids to represent 3D scenes. However, it does not establish non-local dependencies between different hash grids.

To further evaluate the effectiveness of NHA in combined hash grid settings, we conducted experiments on the synthetic NeRF dataset [16]. The results are shown in Figure 8 and Table V. After integrating NHA, hash collisions caused by model compression are alleviated, and the performance of Bi-NeRF is improved, with only a modest increase in model parameters (0.05 M). These findings indicate that our method also achieves strong performance when applied to compressed hybrid hash grid structures.

TABLE V
QUANTITATIVE EVALUATION RESULTS ON SYNTHETIC-NERF DATASETS. THE SYMBOL †INDICATES THAT NHA IS USED. THE BEST METHODS ARE HIGHLIGHTED BY **BOLD**.

Method	PSNR	SSIM	LIPIS	Param
Bi-NeRF-S2	32.01	0.9532	0.0567	0.47 M
Bi-NeRF-B2	32.65	0.9577	0.0491	1.48 M
Bi-NeRF-S2†	32.27	0.9549	0.0521	0.52 M
Bi-NeRF-B2†	**32.81**	**0.9598**	**0.0435**	1.53 M

TABLE VI
ABLATION EXPERIMENTS OF KEY COMPONENTS.

Baseline	3D hash.	MPE	NHA	Time	Param	PSNR	SSIM
✓				16 min	112.77 M	21.83	0.7320
✓	✓			21 min	159.48 M	23.21	0.8435
✓	✓	✓		25 min	159.50 M	24.88	0.8693
✓	✓		✓	41 min	159.51 M	26.41	0.9211
✓	✓	✓	✓	**45 min**	**159.53 M**	**29.96**	**0.9353**

D. Ablation Study

We conduct ablation experiments on medical datasets to study the impact of different components on model performance.

1) Key Components Ablation: We conducted the ablation study to evaluate the impact of key components on the model's reconstruction performance and efficiency. Using a single 4D hash grid as the baseline, we sequentially incorporated additional components: a 3D hash grid for combined representation (3D hash.), Non-local Hash Attention (NHA), and Mask Positional Encoder (MPE). The quantitative results are shown in Table VI.

TABLE VII
ABLATION EXPERIMENTS OF COMBINED HASH GRIDS. THE SYMBOL ‡INDICATES THAT MPE AND NHA ARE USED.

Method	Time	Param	PSNR	SSIM
Baseline	16 min	112.77 M	21.83	0.7320
4D + 4D	26 min	225.51 M	21.97	0.7560
4D + 4D ‡	51 min	225.55 M	23.08	0.8268
4 × 3D	29 min	186.89 M	24.11	0.8673
4 × 3D ‡	72 min	186.97 M	29.78	0.9329
4D + 4 × 3D	40 min	299.54 M	24.78	0.8659
4D + 4 × 3D ‡	87 min	299.73 M	29.76	0.9315
3D + 4D	21 min	159.48 M	23.21	0.8435
3D + 4D †(Ours)	45 min	159.53 M	**29.95**	**0.9353**

TABLE VIII
ABLATION EXPERIMENTS OF ATTENTION.

Method	Time	Param	PSNR	SSIM
Multi-Head Attention	68 min	159.48 M	**30.21**	0.9320
Linear-attention	90 min	159.53 M	29.62	0.9284
Performer	90 min	159.53 M	30.08	0.9322
Ours	45 min	159.53 M	29.95	**0.9353**

In terms of reconstruction performance, the baseline achieves a PSNR score of 21.83 dB. When the 3D hash grid is added for combined representation, the performance improves by 1.38 dB. Further incorporating the MPE results in an additional improvement of 1.67 dB. The inclusion of NHA leads to a significant improvement of 5.08 dB. While NHA noticeably increases the training time, it also leads to a substantial boost in performance.

It is worth noting that the performance improvement from line 2 to line 3 and from line 4 to line 5 in Table VI suggests that MPE is more effective when NHA is present. This is because MPE not only introduces global continuity but also assists NHA in spatial positioning, allowing it to attend based on spatial context rather than solely on raw hash features.

In terms of storage size, the overall number of parameters in the model primarily depends on the size and number of hash grids. Specifically, the addition of the 3D hash grid increases the number of parameters by 46.71 M, while the other components have minimal impact on the total number of parameters. Notably, we use a 159.53 M parameter model to represent approximately 3.2 G of 4D CT image, achieving a compression rate of 20 times while maintaining high fidelity, demonstrating the model's compactness.

Figure 7 more intuitively illustrates the impact of different components on the visual quality of the reconstruction results. It can be observed that both MPE and NHA enhance image clarity in detailed regions (e.g., vessels) and improve the smoothness of homogeneous regions (e.g., liver). More dynamic reconstruction results can be seen in the Supplementary Material.

2) Combined Hash Grid Analysis: We tested various hash grid combinations, including two 4D hash grids (4D + 4D), a combination of four 3D hash grids that are mutually orthogonal in 4D space (4 × 3D: xyz, xyt, xzt, yzt), and a combination of 4D + 4 × 3D, and analyzed their effects on the reconstruction performance. The results are shown in Table VII.

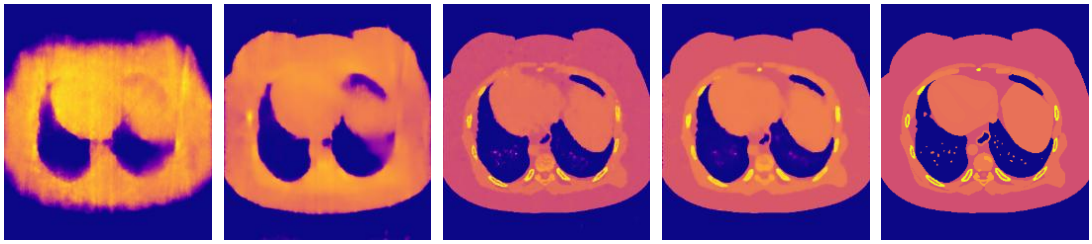
It can be observed that when no 3D hash grid representing the static region (4D + 4D), the reconstruction quality is poor, and even with the use of NHA and MPE, the improvement is limited. This is because the static hash grid more effectively alleviate the temporal redundancy of the static region. The 4 × 3D combination method includes a static hash grid and applies the ideas of decomposition, yielding good results. However, compared to our combination method, it introduces more hash grids, thereby increasing the sequence length of non-local operations and increasing training time. The combination of 4D + 4 × 3D faces similar efficiency problems. While more parameters help capture details, they also increase the risk of overfitting. Considering the number of parameters, reconstruction time and reconstruction quality, the 3D + 4D combination is the best choice.

3) Non-local Hash Attention Analysis: As shown in Figure 3, Figure 7, and Table VI, NHA significantly enhances reconstruction quality. To further evaluate its effectiveness, we replaced NHA with other attention approaches, including Multi-Head Attention [56], Linear Attention [57], and Performer [62], and analyzed their performance within the model. The results are summarized in Table VIII.

It can be seen that different attention approaches provide similar functionality with negligible differences in reconstruction performance, but they result in a notable increase in reconstruction time. This is because both linear $\mathcal{O}(TD^2)$ and quadratic $\mathcal{O}(T^2D)$ complexity approaches involve the dot product, which significantly prolongs training. In addition, recent linear attention approaches, such as Mamba [63], [64] and RWKV [65] are designed for longer sequences. However, this paper only needs to establish non-local dependencies for extremely short sequences (sequence length is 2), rendering such approaches less suitable for our application. In contrast, NHA completely eliminates the dot product and reduces the complexity to $\mathcal{O}(TD)$ without compromising performance.

4) Sparse View and Regularization Analysis: We analyze the reconstruction quality of different methods under sparse views, which is critical for radiation dose reduction studies. Figure 10 presents the reconstruction results for varying sparse numbers of views. Surprisingly, at 40 views, our method outperforms other algorithms. However, at 20 views, our method shows a slight drop in PSNR compared to K-plane. This is because, unlike other methods, our framework does not incorporate regularization or prior knowledge in order to preserve strong generalization across diverse CT scenarios. However, in cases of extremely limited views, regularization can more effectively constrain the solution space and reduce reconstruction ambiguity.

To investigate the effect of regularization on sparse view reconstruction quality, we introduced Total Variation (TV) regularization loss with a weight coefficient of 1e-3, and evaluated its performance under varying numbers of input views. As shown in Figure 9 and Figure 10, when the number of views is extremely sparse (less than 60), TV loss significantly suppresses reconstruction noise. However, as the number of views increases, TV loss tends to introduce over-smoothing, which may hinder the preservation of high-frequency details. In such cases, reducing the weight of the TV loss could improve performance. However, doing so requires



(a) (b) (c) (d) (e)

Fig. 9. Visual analysis of TV loss under different numbers of views. (a) 20 views without TV loss. (b) 20 views with TV loss. (c) 100 views without TV loss. (d) 100 views with TV loss. (e) Ground Truth.

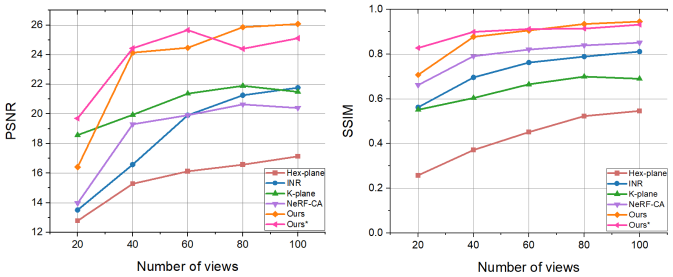


Fig. 10. The number of training views analysis. * indicates using TV loss.

careful tuning to adapt to different CT scenarios, which may be difficult to generalize in practical applications.

Therefore, to balance model generalization with reconstruction performance under sparse view conditions, we recommend applying TV regularization only when the number of input views is extremely limited (e.g., 60 views or fewer).

V. Conclusion and Limitations

In this paper, we propose a novel 4D CT reconstruction method NG-NeRF. We represent the 4D CT scene using the combination of 3D and 4D hash grids, and design non-local hash attention to establish dependencies between different hash grids. This mechanism plays an important role in optimizing feature selection and balancing the hash table load, thereby significantly alleviating the issue of hash collisions and achieving the decoupling of dynamic and static regions. Additionally, we incorporate mask positional encoding to introduce global information, prevent model overfitting to noise, and further mitigate the impact of hash collisions. Our method does not rely on regularization techniques and complex hyperparameter tuning. Experimental results on medical and industrial datasets with different scanning modes demonstrate that the proposed method outperforms existing methods in reconstruction quality, and provides valuable insights for tasks such as dense scene reconstruction and model compression.

Limitations: Although the proposed method substantially improves reconstruction quality, the increased training time remains a noteworthy limitation. Certain clinical scenarios, such as the diagnosis of aortic dissection or acute trauma, impose stringent requirements on reconstruction speed.

The current NHA module nearly doubles the training time, yet it contributes significantly to the improvement of reconstruction quality. Therefore, while continuing to utilize its nonlocal design principles, exploring more computationally efficient alternative architectures, such as nonlocal mean filtering [66], is an important direction. In addition, the highly parallelized differentiable rasterization technique used in 3D Gaussian Splatting (3DGS) [67], [68] has demonstrated a substantial reduction in training time. Extending this representation to 4D CT reconstruction tasks holds promise for achieving even faster training time.

References

[1] J. Liu, Y. Hu, J. Yang, Y. Chen, H. Shu, L. Luo, Q. Feng, Z. Gui, and G. Coatrieux, "3d feature constrained reconstruction for low-dose ct imaging," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 5, pp. 1232–1247, 2016.

[2] Y. Momoki, A. Ichinose, Y. Shigeto, U. Honda, K. Nakamura, and Y. Matsumoto, "Characterization of pulmonary nodules in computed tomography images based on pseudo-labeling using radiology reports," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 5, pp. 2582–2591, 2021.

[3] Q. Zhou, J. Qin, X. Xiang, Y. Tan, and Y. Ren, "Mols-net: Multi-organ and lesion segmentation network based on sequence feature pyramid and attention mechanism for aortic dissection diagnosis," *Knowledge-Based Systems*, vol. 239, p. 107853, 2022.

[4] Y. Zhang, Z. Jiang, Y. Zhang, and L. Ren, "A review on 4d cone-beam ct (4d-cbct) in radiation therapy: Technical advances and clinical applications," *Medical physics*, vol. 51, no. 8, pp. 5164–5180, 2024.

[5] A. W. Reed, H. Kim, R. Anirudh, K. A. Mohan, K. Champley, J. Kang, and S. Jayasuriya, "Dynamic ct reconstruction from limited views with implicit neural representations and parametric motion fields," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2258–2268.

[6] L. A. Feldkamp, L. C. Davis, and J. W. Kress, "Practical cone-beam algorithm," *Josa a*, vol. 1, no. 6, pp. 612–619, 1984.

[7] I. Vergalasova and J. Cai, "A modern review of the uncertainties in volumetric imaging of respiratory-induced target motion in lung radiotherapy," *Medical physics*, vol. 47, no. 10, pp. e988–e1008, 2020.

[8] H. Zhang, J. Ma, Z. Bian, D. Zeng, Q. Feng, and W. Chen, "High quality 4d cone-beam ct reconstruction using motion-compensated total variation regularization," *Physics in Medicine & Biology*, vol. 62, no. 8, p. 3313, 2017.

[9] M. Brehm, P. Paysan, M. Oelhafen, and M. Kachelrieß, "Artifact-resistant motion estimation with a patient-specific artifact model for motion-compensated cone-beam ct," *Medical physics*, vol. 40, no. 10, p. 101913, 2013.

[10] Q. Zhang, Y.-C. Hu, F. Liu, K. Goodman, K. E. Rosenzweig, and G. S. Mageras, "Correction of motion artifacts in cone-beam ct using a patient-specific respiratory motion model," *Medical physics*, vol. 37, no. 6Part1, pp. 2901–2909, 2010.

[11] T. Li, A. Koong, and L. Xing, "Enhanced 4d cone-beam ct with inter-phase motion model," *Medical physics*, vol. 34, no. 9, pp. 3688–3695, 2007.

[12] G. Chee, D. O'Connell, Y. Yang, K. Singhrao, D. Low, and J. Lewis, "Mcsart: an iterative model-based, motion-compensated sart algorithm for cbct reconstruction," *Physics in Medicine & Biology*, vol. 64, no. 9, p. 095013, 2019.

[13] S. Zhi, M. Kachelrieß, F. Pan, and X. Mou, "Cycn-net: A convolutional neural network specialized for 4d cbct images refinement," *IEEE Transactions on Medical Imaging*, vol. 40, no. 11, pp. 3054–3064, 2021.

[14] P. Yang, X. Ge, T. Tsui, X. Liang, Y. Xie, Z. Hu, and T. Niu, "Four-dimensional cone beam ct imaging using a single routine scan via deep learning," *IEEE Transactions on Medical Imaging*, vol. 42, no. 5, pp. 1495–1508, 2022.

[15] Z. Deng, H. Chen, H. Hu, Z. Xu, T. Lyu, Y. Xi, Y. Chen, and J. Zhao, "Rstar: Rotational streak artifact reduction in 4d cbct using separable and circular convolutions," *arXiv preprint arXiv:2403.16361*, 2024.

[16] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[17] K. W. Maas, D. Ruijters, A. Vilanova, and N. Pezzotti, "Nerf-ca: Dynamic reconstruction of x-ray coronary angiography with extremely sparse-views," *arXiv preprint arXiv:2408.16355*, 2024.

[18] Y. Zhang, H.-C. Shao, T. Pan, and T. Mengke, "Dynamic cone-beam ct reconstruction using spatial and temporal implicit neural representation learning (stinr)," *Physics in Medicine & Biology*, vol. 68, no. 4, p. 045005, 2023.

[19] H.-C. Shao, T. Mengke, T. Pan, and Y. Zhang, "Dynamic cbct imaging using prior model-free spatiotemporal implicit neural representation (pmf-stinr)," *Physics in Medicine & Biology*, vol. 69, no. 11, p. 115030, 2024.

[20] T. Müller, A. Evans, C. Schied, and A. Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM transactions on graphics (TOG)*, vol. 41, no. 4, pp. 1–15, 2022.

[21] T. Wu, F. Zhong, A. Tagliasacchi, F. Cole, and C. Oztireli, "D^ 2nerf: Self-supervised decoupling of dynamic and static objects from a monocular video," *Advances in neural information processing systems*, vol. 35, pp. 32 653–32 666, 2022.

[22] R. Zeng, J. A. Fessler, and J. M. Balter, "Estimating 3-d respiratory motion from orbiting views by tomographic image registration," *IEEE transactions on medical imaging*, vol. 26, no. 2, pp. 153–163, 2007.

[23] J. Wang and X. Gu, "Simultaneous motion estimation and image reconstruction (smeir) for 4d cone-beam ct," in *Medical Imaging 2014: Physics of Medical Imaging*, vol. 9033. SPIE, 2014, pp. 762–769.

[24] H. Zhang, J. Ma, Z. Bian, D. Zeng, Q. Feng, and W. Chen, "High quality 4d cone-beam ct reconstruction using motion-compensated total variation regularization," *Physics in Medicine & Biology*, vol. 62, no. 8, p. 3313, 2017.

[25] Z. Qi and G.-H. Chen, "Extraction of tumor motion trajectories using piccs-4dcbct: a validation study," *Medical physics*, vol. 38, no. 10, pp. 5530–5538, 2011.

[26] J. He, Y. Yang, Y. Wang, D. Zeng, Z. Bian, H. Zhang, J. Sun, Z. Xu, and J. Ma, "Optimizing a parameterized plug-and-play admm for iterative low-dose ct reconstruction," *IEEE transactions on medical imaging*, vol. 38, no. 2, pp. 371–382, 2018.

[27] Y. Li, K. Li, C. Zhang, J. Montoya, and 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*, vol. 38, no. 10, pp. 2469–2481, 2019.

[28] X. Zhu, J. Zhou, L. You, X. Yang, J. Chang, J. J. Zhang, and D. Zeng, "Dfie3d: 3d-aware disentangled face inversion and editing via facial-contrastive learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 9, pp. 8310–8326, 2024.

[29] Y. Li, Q. Hu, Z. Ouyang, and S. Shen, "Neural reflectance decomposition under dynamic point light," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 4, pp. 2195–2208, 2024.

[30] Z. Sheng, F. Liu, M. Liu, F. Zheng, and L. Nie, "Open-set synthesis for free-viewpoint human body reenactment of novel poses," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 12, pp. 12 676–12 691, 2024.

[31] T. Zhang, L. Zhang, F. Zhang, S. Zhao, and Y. Zhou, "I-dacs: Always maintaining consistency between poses and the field for radiance field construction without pose prior," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2024.

[32] J. Yang, M. Pavone, and 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.

[33] M. Bonotto, L. Sarrocco, D. Evangelista, M. Imperoli, and A. Pretto, "Combinerf: A combination of regularization techniques for few-shot neural radiance field view synthesis," in *2024 International Conference on 3D Vision (3DV)*. IEEE, 2024, pp. 641–650.

[34] S. Guo, Q. Wang, Y. Gao, R. Xie, L. Li, F. Zhu, and L. Song, "Depth-guided robust point cloud fusion nerf for sparse input views," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 9, pp. 8093–8106, 2024.

[35] A. Chen, Z. Xu, A. Geiger, J. Yu, and H. Su, "Tensorf: Tensorial radiance fields," in *European conference on computer vision*. Springer, 2022, pp. 333–350.

[36] J. Ding, Y. He, B. Yuan, Z. Yuan, P. Zhou, J. Yu, and X. Lou, "Ray reordering for hardware-accelerated neural volume rendering," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 11, pp. 11 413–11 422, 2024.

[37] S. Shin and J. Park, "Binary radiance fields," *Advances in neural information processing systems*, vol. 36, 2024.

[38] Y. Chen, Q. Wu, M. Harandi, and J. Cai, "How far can we compress instant-ngp-based nerf?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 20 321–20 330.

[39] Z. Zheng, F. Lu, W. Xue, G. Chen, and C. Jiang, "Lidar4d: Dynamic neural fields for novel space-time view lidar synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 5145–5154.

[40] C. Chen, S. Huang, X. Chen, G. Chen, X. Han, K. Zhang, and M. Gong, "Ct4d: Consistent text-to-4d generation with animatable meshes," *arXiv preprint arXiv:2408.08342*, 2024.

[41] A. Pumarola, E. Corona, G. Pons-Moll, and F. Moreno-Noguer, "D-nerf: Neural radiance fields for dynamic scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 318–10 327.

[42] K. Park, U. Sinha, J. T. Barron, S. Bouaziz, D. B. Goldman, S. M. Seitz, and R. Martin-Brualla, "Nerfies: Deformable neural radiance fields," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5865–5874.

[43] L. Song, A. Chen, Z. Li, Z. Chen, L. Chen, J. Yuan, Y. Xu, and A. Geiger, "Nerfplayer: A streamable dynamic scene representation with decomposed neural radiance fields," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 5, pp. 2732–2742, 2023.

[44] K. Park, U. Sinha, P. Hedman, J. T. Barron, S. Bouaziz, D. B. Goldman, R. Martin-Brualla, and S. M. Seitz, "Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields," *arXiv preprint arXiv:2106.13228*, 2021.

[45] S. Park, M. Son, S. Jang, Y. C. Ahn, J.-Y. Kim, and N. Kang, "Temporal interpolation is all you need for dynamic neural radiance fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 4212–4221.

[46] A. Cao and J. Johnson, "Hexplane: A fast representation for dynamic scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 130–141.

[47] S. Fridovich-Keil, G. Meanti, F. R. Warburg, B. Recht, and A. Kanazawa, "K-planes: Explicit radiance fields in space, time, and appearance," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12 479–12 488.

[48] Y. Cai, J. Wang, A. Yuille, Z. Zhou, and A. Wang, "Structure-aware sparse-view x-ray 3d reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 11 174–11 183.

[49] D. Rückert, Y. Wang, R. Li, R. Idoughi, and W. Heidrich, "Neat: Neural adaptive tomography," *ACM Transactions on Graphics (TOG)*, vol. 41, no. 4, pp. 1–13, 2022.

[50] R. Zha, Y. Zhang, and 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.

[51] Q. Zhou, Y. Ye, and Z. Cai, "Spatiotemporal-aware neural fields for dynamic ct reconstruction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 10, 2025, pp. 10 834–10 842.

[52] Q. Zhou, G. Lu, Yunfan, and Z. Cai, "Deblurtomo: Self-supervised computed tomography reconstruction from blurry images," *Computers, Materials & Continua*, p. pages. [Online]. Available: http://www.techscience.com/cmc/online/detail/23457

[53] A. C. Kak and M. Slaney, *Principles of computerized tomographic imaging*. SIAM, 2001.

[54] M. Teschner, B. Heidelberger, M. Müller, D. Pomerantes, and M. H. Gross, "Optimized spatial hashing for collision detection of deformable objects." in *Vmv*, vol. 3, 2003, pp. 47–54.

[55] S. Zhai, W. Talbott, N. Srivastava, C. Huang, H. Goh, R. Zhang, and J. Susskind, "An attention free transformer," *arXiv preprint arXiv:2105.14103*, 2021.

[56] A. Waswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017.
[57] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret, "Transformers are rnns: Fast autoregressive transformers with linear attention," in *International conference on machine learning*. PMLR, 2020, pp. 5156–5165.
[58] L. Xu, Y. Xiangli, S. Peng, X. Pan, N. Zhao, C. Theobalt, B. Dai, and D. Lin, "Grid-guided neural radiance fields for large urban scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 8296–8306.
[59] W. P. Segars, M. Mahesh, T. J. Beck, E. C. Frey, and B. M. Tsui, "Realistic ct simulation using the 4d xcat phantom," *Medical physics*, vol. 35, no. 8, pp. 3800–3808, 2008.
[60] G. D. Hugo, E. Weiss, W. C. Sleeman, S. Balik, P. J. Keall, J. Lu, and J. F. Williamson, "A longitudinal four-dimensional computed tomography and cone beam computed tomography dataset for image-guided radiation therapy research in lung cancer," *Medical physics*, vol. 44, no. 2, pp. 762–771, 2017.
[61] Z. Deng, W. Zhang, K. Chen, Y. Zhou, J. Tian, G. Quan, and J. Zhao, "Tt u-net: Temporal transformer u-net for motion artifact reduction using pad (pseudo all-phase clinical-dataset) in cardiac ct," *IEEE Transactions on Medical Imaging*, 2023.
[62] K. Choromanski, V. Likhosherstov, D. Dohan, X. Song, A. Gane, T. Sarlos, P. Hawkins, J. Davis, A. Mohiuddin, L. Kaiser *et al.*, "Rethinking attention with performers," *arXiv preprint arXiv:2009.14794*, 2020.
[63] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.
[64] T. Dao and A. Gu, "Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality," in *International Conference on Machine Learning (ICML)*, 2024.
[65] B. Peng, E. Alcaide, Q. Anthony, A. Albalak, S. Arcadinho, S. Biderman, H. Cao, X. Cheng, M. Chung, M. Grella *et al.*, "Rwkv: Reinventing rnns for the transformer era," *arXiv preprint arXiv:2305.13048*, 2023.
[66] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognitio*, vol. 2. Ieee, 2005, pp. 60–65.
[67] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering." *ACM Trans. Graph.*, vol. 42, no. 4, pp. 139–1, 2023.
[68] R. Zha, T. J. Lin, Y. Cai, J. Cao, Y. Zhang, and H. Li, "R^2-gaussian: Rectifying radiative gaussian splatting for tomographic reconstruction," *arXiv preprint arXiv:2405.20693*, 2024.

Qingyang Zhou received his B.E. and M.S. degrees from the Central South University of Forestry and Technology, in 2018 and 2022, respectively. He is currently pursuing a Ph.D. degree in the College of Computer Science and Technology at the National University of Defense Technology. His research interests are in the domain of CT reconstruction, neural radiation field, medical image segmentation and object detection.

Yunfan Ye is an Assistant Professor in School of Design, Hunan University (HNU), China. He earned his PhD degree in December 2023 at the National University of Defense Technology, under the supervision of Prof. Zhiping Cai and Prof. Kai Xu in iGrape Lab. He got his M.S. degree in Computer Science in 2019 from the Stevens Institute of Technology and his B.E. degree in Computer Science in 2017 from Xiamen University, China.

Zhihuang Liu received his B.E. and M.S. degrees from the College of Computer and Data Science, Fuzhou University, in 2020 and 2023, respectively. He is currently pursuing a Ph.D. degree in the College of Computer Science and Technology at the National University of Defense Technology. His research interests include medical privacy protection, applied cryptography, and privacy in generative AI.

Chang Liu received the B.Eng. degree in Computer Science and Technology from Lanzhou University, China, in 2022. He is currently pursuing his doctoral studies at the National University of Defense Technology (NUDT), focusing on medical image processing, deepfake video detection, and video understanding.

Zhiping Cai received the B.Eng., M.A.Sc., and Ph.D. degrees in computer science and technology from the National University of Defense Technology (NUDT), China, in 1996, 2002, and 2005, respectively. He is a full professor in the College of Computer, NUDT. His current research interests include artificial intelligence, network security and big data. He is a senior member of the CCF and a member of the IEEE.