Spatiotemporal-Aware Neural Fields for Dynamic CT Reconstruction

Qingyang Zhou[1], **Yunfan Ye**[2*], **Zhiping Cai**[1]

[1]College of Computer Science and Technology, National University of Defense Technology, Changsha, China
[2]School of Design, Hunan University, Changsha, China

Abstract

We propose a dynamic Computed Tomography (CT) reconstruction framework called *STNF4D* (SpatioTemporal-aware Neural Fields). First, we represent the 4D scene using four orthogonal volumes and compress these volumes into more compact hash grids. Compared to the plane decomposition method, this method enhances the model's capacity while keeping the representation compact and efficient. However, in densely predicted high-resolution dynamic CT scenes, the lack of constraints and hash conflicts in the hash grid features lead to obvious dot-like artifact and blurring in the reconstructed images. To address these issues, we propose the Spatiotemporal Transformer (ST-Former) that guides the model in selecting and optimizing features by sensing the spatiotemporal information in different hash grids, significantly improving the quality of reconstructed images. We conducted experiments on medical and industrial datasets covering various motion types, sampling modes, and reconstruction resolutions. Experimental results show that our method outperforms the second-best by 5.99 dB and 4.11 dB in medical and industrial scenes, respectively.

Code — https://qingyangzhou69.github.io/STNF4D

Introduction

Traditional Computed Tomography (CT) reconstruction technology leverages the penetrating properties of X-rays to reveal detailed internal structures of objects (Kak and Slaney 2001). In scenes where the scanned object is moving (4D-CT), 3D-CT reconstruction technology struggles to capture the dynamic process, which has various applications ranging from medical radiation therapy (Jaffray et al. 2002; Zhang et al. 2024) to industrial material analysis (Reed et al. 2021).

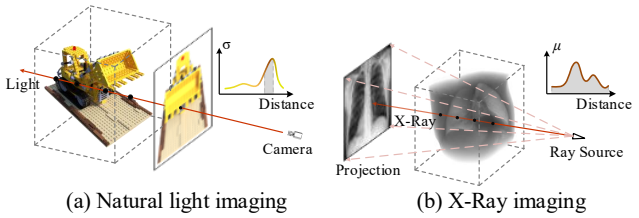
The process of medical 4D-CT scanning generally applies motion sensors to capture motion state (phase) signals and collect projections only at specific phases. The vanilla 4D reconstruction method uses the projections of each phase for independent reconstruction (gated reconstruction). However, this method will cause streak artifacts under the limited number of projection acquisitions (Vergalasova and Cai 2020). Existing learning-based methods typically utilize Convolutional Neural Networks (CNNs) to remove streak artifacts in the image domain (Zhi et al. 2021; Yang et al. 2022; Deng et al. 2024), achieving promising results in the field of medical 4D-CT reconstruction. Nevertheless, these methods require a large amount of training data, and due to the inherent limitations of CNNs and computational constraints, these methods encounter challenges in capturing 4D spatiotemporal information and generalizing across different scenes and is difficult to apply to scenes where the motion state is unknown (e.g. industrial scene).

Compared to medical scenes, it becomes even more challenging in industrial applications to capture aperiodic object motions. Existing methods for addressing such aperiodic motion (Reed et al. 2021; Mohan et al. 2015) typically focus on low-resolution reconstruction scenes.

Recently, the success of NeRF (Mildenhall et al. 2021; Chen et al. 2023; Liao et al. 2024) in natural scene reconstruction has provided a feasible solution for CT reconstruction (Zha, Zhang, and Li 2022; Rückert et al. 2022; Cai et al. 2024). However, existing NeRF-based 4D-CT reconstruction methods (Reed et al. 2021; Zhang et al. 2023; Shao et al. 2024) only support dynamic CT reconstruction at low resolutions. There are two possible reasons. First, due to the nature of X-rays to penetrate objects, they usually bring more information than natural light, as shown in Figure 1. CT scenes need to pay more attention to the entire reconstruction space (with a higher rank), especially in 4D high-resolution scenes, thereby imposing greater demands on model representation capabilities. Secondly, although existing grid-based (Fang et al. 2022; Müller et al. 2022) and plane-based (Cao and Johnson 2023; Fridovich-Keil et al. 2023; Shao et al. 2023) methods can flexibly improve model capacity, such discrete features are not fully utilized with the lack of constraints. Simply expanding the model scale also brings limited improvement.

To address the above problems, we propose a new framework, STNF4D. Inspired by plane decomposition (Fridovich-Keil et al. 2023; Cao and Johnson 2023), we decompose the 4D scene into four orthogonal volumes that map to hash grids. This approach aims at a more compact and efficient 4D representation. However, in densely predicted high-resolution dynamic CT scenes, the lack of constraints and hash conflicts in the hash grid features result in obvious dot-like artifacts and blurring in the recon-

*Corresponding author



(a) Natural light imaging (b) X-Ray imaging

Figure 1: (a) Natural scene reconstruction focuses on the object surface. (b) CT scene reconstruction focuses on the entire spatial volume.

structed images. To overcome this issue, we proposed a Spatiotemporal Transformer (ST-Former), which performs self-attention on the corresponding spatiotemporal features of the sampling points in different hash grids, guiding the model in feature selection and significantly improving the quality of the reconstructed images. We conduct experiments on high-resolution medical and industrial datasets to evaluate the reconstruction performance across various scenes and sampling modes. Our method consistently outperforms the existing approaches by a significant margin in all datasets. Our contributions can be summarized as follows:

- A self-supervised framework that achieves high-quality 4D-CT reconstruction across various scenes.
- Several technical designs, including the 4D hash grid representation and spatiotemporal transformer, ensure the reconstruction quality and reduce hash conflicts.
- Extensive experiments on medical and industrial datasets demonstrate the superiority of our method under various scenes and sampling modes.

Related Work

4D-CT Reconstruction

Traditional 4D-CT reconstruction methods typically use motion modeling to create deformation vector fields (DVFs) that compensate for object motion and reduce reconstruction artifacts. DVFs can be obtained through non-rigid registration between 4D-CT images of different phases (Brehm et al. 2013; Zhang et al. 2010; Li, Koong, and Xing 2007), or directly from projection data (Fridovich-Keil et al. 2023; Zeng, Fessler, and Balter 2007). However, registration techniques can introduce errors, and optimizing DVFs is challenging. Many methods (Wang and Gu 2013; Zhang et al. 2017) address this by incorporating multiple regularization terms based on prior knowledge into the reconstruction process. Nonetheless, excessive regularization can limit the diversity of motion representation and impair the algorithm's adaptability to various motion and sampling modes (Chee et al. 2019).

Learning-based methods for reconstruction often focus on reducing streak artifacts in the image domain by training CNNs on large artifact-free datasets (Qi and Chen 2011). However, this approach requires extensive datasets without artifacts and cannot be applied to unknown scenes. Furthermore, existing methods capture relevant information between 2D CT slices at different phases using 3D convolution (Deng et al. 2024), 2D+t convolution (Zhi et al. 2021), or Transformer (Deng et al. 2023). However, due to limitations in computing power, learning-based methods often disregard the spatial information of the CT volume within the same time sequence, thereby limiting reconstruction performance.

Although widely adopted, previous methods have struggled to meet the demands of generalization across various scenes and high-quality 4D-CT reconstruction. We believe that the community still needs a method that simultaneously satisfies both demands.

Neural Rendering for Dynamic Scene

Neural Radiance Fields (NeRF) for dynamic scene reconstruction is a significant branch of NeRF research. Generally, dynamic NeRF can be represented in two ways. The first decouples the dynamic scene into a canonical space and a time-varying deformation field (Pumarola et al. 2021; Park et al. 2021; Song et al. 2023). This method offers high flexibility, but usually relies on implicit neural representations and exhibits low convergence and inference speeds. The second approach represents the 4D scene using voxel grids (Fang et al. 2022; Müller et al. 2022) or planes (Shao et al. 2023; Cao and Johnson 2023; Fridovich-Keil et al. 2023). This method models the space more explicitly, can flexibly improve the capacity of the model, and improves both training speed and rendering accuracy. However, the lack of sufficient constraints on discrete features leads to insufficient feature utilization.

The K-plane (Fridovich-Keil et al. 2023) and Hexplane (Cao and Johnson 2023) used six planes to represent 4D scenes. Inspired by this, we decompose the 4D scene into 3D volumes and establish dependencies between the spatiotemporal features in different volumes to improve feature utilization.

Neural Rendering for Dynamic CT Reconstruction

Recently, many studies have developed CT reconstruction algorithms based on NeRF (Rückert et al. 2022; Cai et al. 2024; Zha, Zhang, and Li 2022), demonstrating the potential of NeRF in static CT. However, in the field of dynamic reconstruction, great challenges are still faced. Existing methods (Reed et al. 2021; Zhang et al. 2023; Shao et al. 2024) model dynamic CT scenes as a combination of a static field using implicit neural representation and a dynamic field that evolves over time, similar to the decoupling methods of NeRF. While this method leverages the flexibility of implicit neural representation to better accommodate complex nonlinear motions, it overlooks the model's representation capacity and is limited to lower-resolution reconstructions.

In this paper, we improve the model's representation ability by introducing 4D hash grid representation and spatiotemporal Transformer.

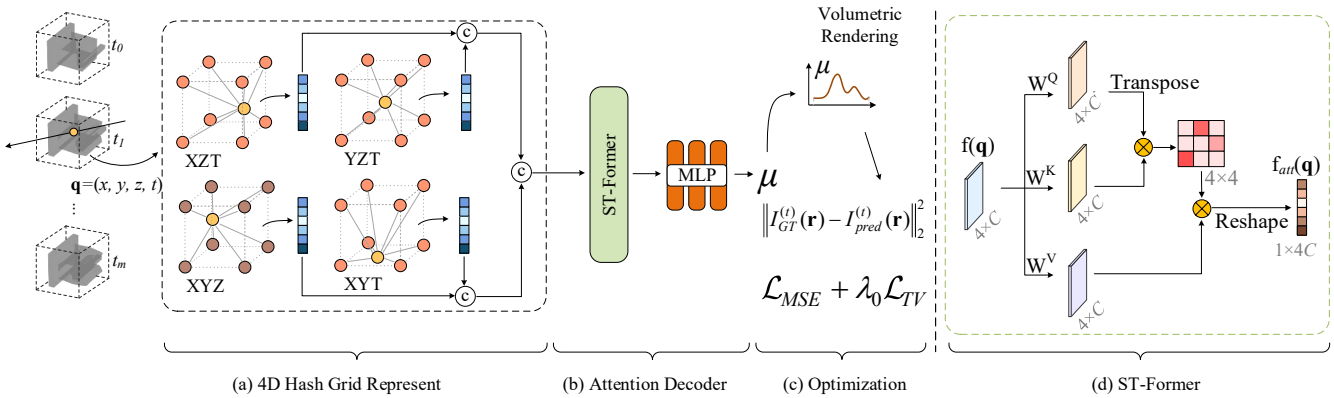


Figure 2: Method overview. (a) We decompose the dynamic scene into four orthogonal volumes that map to hash grids. The coordinate **q** is projected into the four volumes respectively to obtain the spatiotemporal features **f(q)**. (b) Then, **f(q)** is input into (d) ST-Former to establish the spatiotemporal feature dependencies, obtain the attention features $\mathbf{f}_{att}$**(q)**, and use a tiny MLP to decode $\mathbf{f}_{att}$**(q)** to obtain the predicted attenuation coefficient μ. (c) We follow Beer's law for volume rendering, obtain the predicted projection values, and use the TV regularization and MSE loss to optimize the model.

Method

In this section, we first introduce the problem formulation of the 4D-CT reconstruction inspired by NeRF. Then, we provide a detailed description of the STNF4D framework including the 4D hash grid representation and spatiotemporal transformer.

Problem Formulation

In natural scenes, NeRF usually learns to map from 3D coordinates $\mathbf{x} \in \mathbb{R}^3$ and viewing direction $\mathbf{d} \in \mathbb{R}^2$ to density $\sigma \in \mathbb{R}$ and color $\mathbf{c} \in \mathbb{R}^3$. In dynamic natural scenes, the time dimension $t \in \mathbb{R}$ is extended, and NeRF can be expressed as:

$$F_\Theta(\mathbf{x}, \mathbf{d}, t) \rightarrow (\mathbf{c}, \sigma). \tag{1}$$

Unlike natural scenes, there is no viewing direction and color information in CT scenes. The imaging principle is shown in Figure 1. In cone beam CT, the ray source emits cone beam X-rays, which attenuate energy after passing through the scanned object. The flat-panel detector detects the attenuated rays energy and generates a projection. According to Beer's law (Kak and Slaney 2001), the attenuation of the rays is related to the object's thickness and its attenuation coefficient. The concept of attenuation coefficient is similar to density. The rendering function can be expressed as:

$$I_{GT}^{(t)}(\mathbf{r}) = I_0 \exp(-\int_{h_n}^{h_f} \mu(\mathbf{r}^p(h), t) dh), \tag{2}$$

where $I_{GT}^{(t)}(\mathbf{r})$ is the energy intensity of ray $\mathbf{r}$ after attenuation, and I_0 represents the bright field value (ray source intensity) detected in the absence of a target. $\mu(\mathbf{r}(h), t)$ represents the attenuation coefficient at the position of ray $\mathbf{r}(h)$ and time t. Among them, $\mathbf{r}(h) = \mathbf{o} + h\mathbf{d}$, $\mathbf{o}$ is the coordinate of the ray source, and $\mathbf{d}$ is the direction vector from the ray source to the detector pixel. h_f and h_n represent the far and near ends of the reconstruction space on the ray, respectively. The ray source and detector rotate around the scanned object to obtain multi-view projections. Our goal is to learn a mapping from coordinates $\mathbf{x}$ and time t to attenuation coefficients through multiple view projection images. The neural attenuation field can be expressed as:

$$F_\Phi(\mathbf{x}, t) \rightarrow \mu, \tag{3}$$

where F_Φ represents a mapping function with learnable parameters Φ. Discretizing Eq.2, we can get the predicted projection:

$$I_{pred}^{(t)}(\mathbf{r}) = I_0 \exp(-\sum_{i=0}^{N} \mu(\mathbf{r}^p(h_i), t)\delta_i), \tag{4}$$

where N represents the number of ray sampling points, while δ_i denotes the distance between consecutive sampling points i and $i+1$. Unlike natural scenes, we employ layered sampling to ensure that the sampling points can cover the entire reconstruction space. It's important to note that in CT reconstruction, the primary task is predicting attenuation coefficients at specific coordinates rather than novel view synthesis. In addition, t represents the motion state in medical 4D-CT reconstruction (see Figure 4 for details).

4D Hash Grid Representation

Compared to natural scenes, which emphasize surface information, CT scenes focus on the entire reconstruction space, resulting in a higher rank, especially in 4D high-resolution medical imaging. This characteristic imposes greater demands on the model's representation capability.

As shown in Figure 2-a, to enhance the model's representation capability, we extend the decomposition method by decomposing the 4D scene into four orthogonal volumes, which are then compressed into the hash grids through hash mapping (Müller et al. 2022). The hash grids are divided into a spatial grid $\mathbf{G}_{XYZ}$ and spatiotemporal grids $\mathbf{G}_{XYT}$,

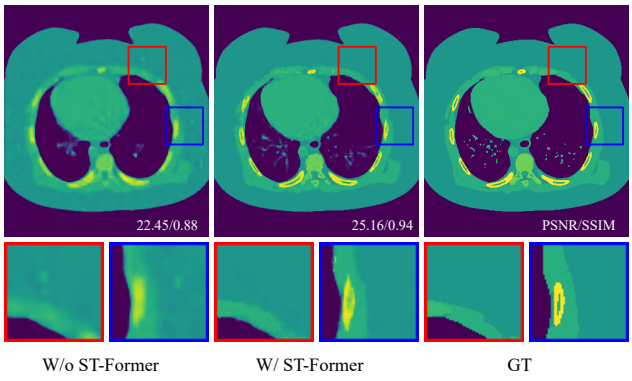


Figure 3: Compare the reconstruction results with and without the ST-Former. It can be seen that SF-Former greatly alleviates dot-like artifacts and blur.

$\mathbf{G}_{YZT}$ and $\mathbf{G}_{XZT}$. Assuming that the temporal and spatial resolutions are M, the size of each grid is $M^3 \times C$, where C represents the length of the feature. For simplicity, we ignore multi-scale resolution. Let the coordinate $\mathbf{q} = (x, y, z, t)$, we normalize it and project it into the hash grid, and we can get the feature:

$$\mathbf{f}(\mathbf{q})_s = P(\mathbf{G}_s, \mathbf{q}), \tag{5}$$

where P is the projection function, and s represents the hash grid index. This projection operation is repeated for each $s \in S$ grid and concatenated grid features as follows:

$$\mathbf{f}(\mathbf{q}) = \bigcup_{s \in S} \mathbf{f}(\mathbf{q})_s, \mathbf{f}(\mathbf{q}) \in \mathbb{R}^{4 \times C}. \tag{6}$$

Different from the K-plane (Fridovich-Keil et al. 2023), which uses the Hadamard product for features in different planes, the Hadamard product can easily cause spatial local signals to shift due to hash conflicts. Therefore, we concatenate features from different hash tables.

Spatiotemporal Transformer

There are hash conflicts in the hash mapping process. Instant-ngp (Müller et al. 2022) addresses this issue by using a tiny Multilayer Perceptron (MLP) to mitigate the effects of these collisions implicitly. However, the dense prediction nature of dynamic CT scenes exacerbates the impact of hash collisions, which is difficult for MLP to handle. Roughly increasing the size of the hash table can alleviate this problem to a certain extent, but it is easy to cause large memory overhead. Additionally, the grid features lack constraints, resulting in low feature utilization and feature redundancy.

To address these challenges, we propose a spatiotemporal Transformer (ST-Former) that enhances feature selection by perceiving the dependency relationships across different hash grids. This approach helps avoid features with hash conflicts and increases feature utilization. As shown in Figure 2-d, let the concatenated feature of the sampling point $\mathbf{q}$ across multiple hash grids be $\mathbf{f}(\mathbf{q}) \in \mathbb{R}^{4 \times C}$. First, we use the linear layer to extract the query, key, and value components from $\mathbf{f}(\mathbf{q})$.

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

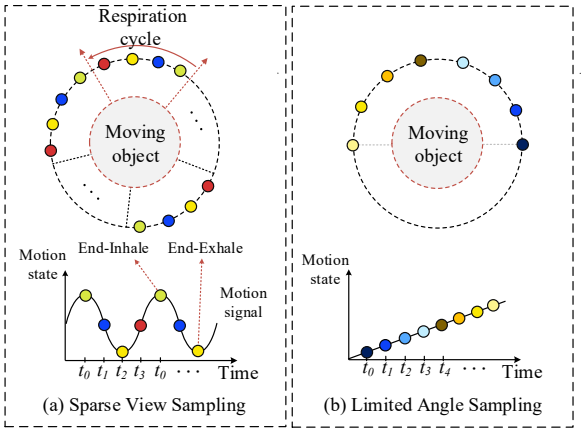


Figure 4: Schematic diagram of sampling views and motion state. (a) Sparse view sampling mode for medical images. Due to the periodicity of motion and phase-gating technology, sampling can be performed from multiple angles in the same motion state (phase). (b) Limited angle sampling mode for general object motion. Due to the unpredictable motion, sampling can only be performed from limited angles in the same motion state.

where W^{Q}, W^{K} and $\mathrm{W}^{\mathrm{V}} \in \mathbb{R}^{C \times C}$. The final output features can be expressed as:

$$\mathbf{f}_{att}(\mathbf{q}) = softmax(\frac{\mathrm{K}^{\top}\mathrm{Q}}{\sqrt{C}})\mathrm{V} + \mathbf{f}(\mathbf{q}). \tag{8}$$

We perform self-attention on the spatiotemporal features of a single sample point in different hash grids. Considering the large number of sampling points, we did not use multi-head attention or more complex attention structures. Nevertheless, the results achieved with this simple approach are remarkable (as shown in Figure 3), and it also offers valuable insights into the issue of hash collisions in 3D reconstruction based on hash encoders.

Optimization

We optimize the model by using the mean square error between the projections calculated by Eq.3 and the measured projections, as well as the total variation (TV) regularization.

$$\mathcal{L}_{total} = \sum\nolimits_{\mathbf{r} \in \mathcal{B}} \left\| I^{(t)}_{pred}(\mathbf{r}) - I^{(t)}_{GT}(\mathbf{r}) \right\|^2_2 + \lambda_0 \mathcal{L}_{TV}(\mathbf{C}), \tag{9}$$

where $\mathcal{B}$ represents the number of rays in a batchsize. We directly compute the TV loss on the predicted attenuation coefficients. To maintain generalizability across various types of motion, we refrained from imposing constraints on object motion.

Experiment

Datasets

Medical dataset Medical 4D-CT usually has a specific sampling mode. Due to the periodicity of breathing, gating technology is used to scan and generate projection images only in a specific motion state (phase). The relationship

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
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
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
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
Ours	**25.16**	**0.9397**	**28.10**	**0.9072**	**31.18**	**0.9261**	**28.79**	**0.9154**	**29.8**	**0.9156**	**28.60**	**0.9208**

Table 1: Quantitative comparison on the medical dataset. We **bold** the best results and underline the second-best.

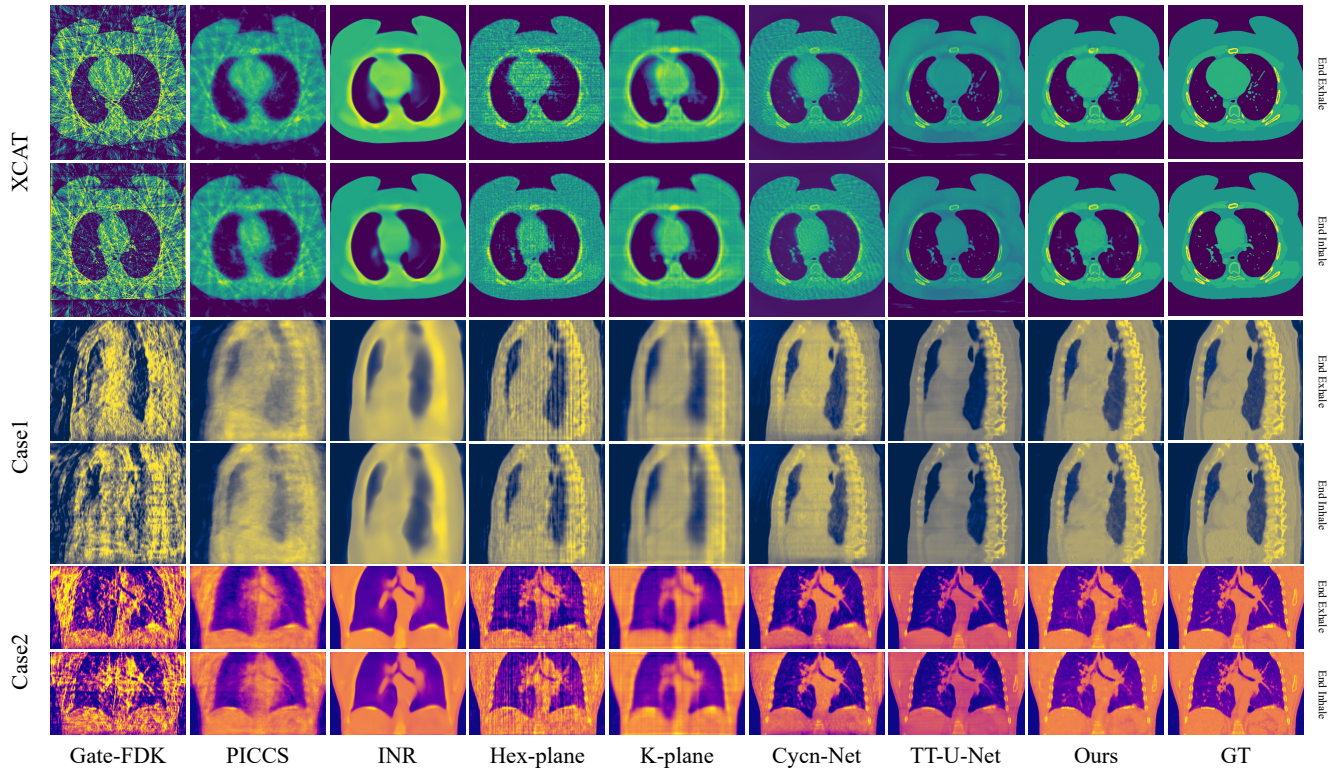


Figure 5: Visualization results of the medical dataset. We assign different colors to different cases and show the reconstruction results of the End-Inhale and End-Exhale for each case. The display window is set to [-1000, 500] HU for reconstructed images.

between projection acquisition and the respiratory phase is shown in Figure 4-a. We collected 4D extended cardiac-torso (XCAT) (Segars et al. 2008) phantom and real patient 4D-CT images from 4D-Lung Cancer Imaging Archive (TCIA) (Hugo et al. 2017), which were consistent with (Zhi et al. 2021). Each 4D-CT image has an XY-axis resolution of 512 × 512, with the Z-axis resolution varying between 70 and 160, and contains 10 CT volumes, representing 10 respiratory phases. We followed the scanning mode of Varian Medical System (Palo Alto, CA) and synthesized 100 projections within 360°, with each phase containing 10 projections at sparse viewing angles.

Industrial dataset Due to the singleness of medical CT scanning modes and the predictability of motion states, we further explore different scanning modes and more general object motion. We collected the dataset (Reed et al. 2021) of aluminum products' damage and evolution under various external forces, which provides valuable information for material performance and safety. Each 4D-CT image set in this dataset has a resolution of 256^3 and contains 60 CT volumes, representing 60 unpredictable motion states. We synthesized 240 projection images within a range of 180°, and each motion state contains projections at 4 different angles. Unlike the medical dataset, the projections of each motion state in the industrial dataset have limited viewing angles, which greatly increases the difficulty of reconstruction, as shown in Figure 4-b.

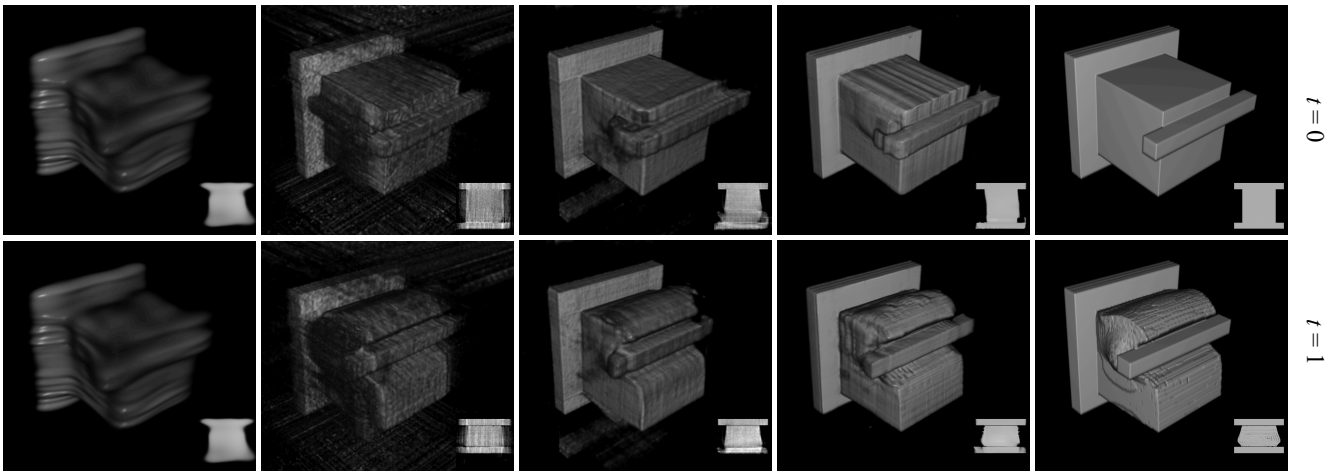


Figure 6: Visualization results of industrial dataset. We show the rendering and tomographic slice results (bottom right) at the beginning ($t = 0$) and end ($t = 1$) of the aluminum product deformation.

Method	Alum 1		Alum 2		Alum 3		Alum 4		Alum 5		Average	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
INR	22.73	0.9379	19.71	0.9066	21.56	0.9100	23.12	0.9457	20.22	0.9155	21.46	0.9231
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
Ours	**26.21**	**0.9795**	**25.59**	**0.9786**	**27.06**	**0.9800**	**27.21**	**0.9824**	**21.77**	**0.9605**	**25.57**	**0.9762**

Table 2: Quantitative comparison on the industrial dataset. We **bold** the best results and underline the second-best.

Implementation Details

Our method is implemented based on Pytorch, and all experiments are completed on a single RTX 4090 GPU. The model is optimized using the Adam optimizer. The initial learning rate is set to 1e-2, and the learning rate is reduced to 1e-6 using the cosine decay strategy. The number of iterations for each scene is 40k, and the number of rays per batch is 1024. In the medical dataset, the number of sampling points per ray is 768; in the industrial dataset, the number of sampling points per ray is 320. We use Peak Signal-to-Noise ratio (PSNR) and Structural Similarity (SSIM) to evaluate the reconstruction quality of the reconstructed images quantitatively. More details can be found in the Supplementary Material.

Comparison

Results on medical dataset To verify the effectiveness of the proposed method, we compared it with traditional methods (Gate-FDK (Feldkamp, Davis, and Kress 1984), PICCS (Qi and Chen 2011)) and NeRF-based methods (K-Plane (Fridovich-Keil et al. 2023), Hex-Plane (Cao and Johnson 2023) and INR (Reed et al. 2021)). In addition, we compared two learning-based 4D-CT reconstruction methods(CycN-Net (Zhi et al. 2021) and TT-U-Net (Deng et al. 2023)), where CycN-Net collected 17 high-quality 4D-CT images from TCIA (Hugo et al. 2017) for training, and TT-U-Net training data was consistent with CycN-Net.

The quantitative results of the medical dataset experiments are shown in Table 1. Compared with the NeRF-based SOTA method K-plane, our method has a PSNR and SSIM that are 5.99 dB and 26.42% higher, respectively. Compared to the best learning-based method (TT-U-Net), our method achieves a PSNR and SSIM increase of 9.49 dB and 9.95%, respectively. It can be seen that the proposed method is far superior to existing 4D-CT reconstruction methods.

Figure 5 shows the visual results of the End-Inhale and End-Exhale phases for 3 cases from the XCAT phantom and TCIA. It can be observed that the traditional method, Gate-FDK, exhibits significant streak artifacts under sparse views, while PICCS uses regularization to smooth the images, but they remain relatively blurry. In the NeRF-based methods, INR lacks sufficient model capacity, resulting in the loss of high-frequency details. Hex-plane and K-plane have grid-like artifacts due to the lack of feature constraints. The learning-based method performs better in overall vision, but due to the difficulty of convolutional neural networks in perceiving 4D information, there are still challenges in artifact removal. More importantly, it cannot be applied to unknown datasets.

Results on industrial dataset To further verify the versatility of the proposed method, we conducted experiments on industrial datasets. Compared with medical datasets, industrial datasets have less spatial high-frequency information

Baseline	4D HGR	ST-Former	PSNR	SSIM
✓			17.13	0.5458
	✓		23.81	0.8530
	✓	✓	28.61	0.9208

Table 3: Key components ablation.

but higher temporal resolution. We compared NeRF-based methods with strong generalization ability and eliminated traditional reconstruction methods and learning-based methods with limited generalization ability. It can be seen that our method can still achieve the best results under different motion types and sampling modes. Compared with the second-based method, PSNR and SSIM are improved by 4.26 dB and 3.23%, respectively.

Figure 6 shows the visual reconstruction results for the initial and final deformation states of an aluminum product under stress. It can be seen that our method has the clearest rendering and slice results. However, INR struggles with sensitivity to temporal changes in high temporal resolution scenes, and it is difficult to form dynamic reconstruction. Hex-plane and K-plane perform slightly better. More visualization results and dynamic results of medical and industrial datasets can be found in the Supplementary Materials.

Ablation Study

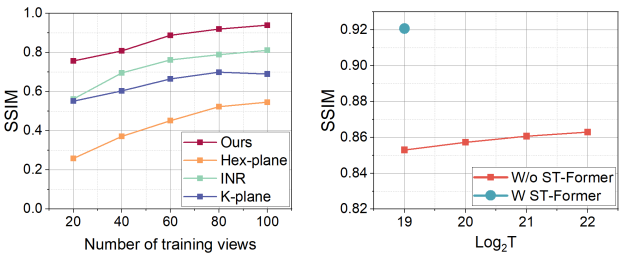
We conduct ablation experiments on challenging medical datasets to further validate the impact of the key components of the proposed method.

Key components ablation We use the six-plane decomposition method Hex-plane as our baseline and add 4D hash grid representation (4D HGR) and ST-Former to perform ablation studies. The results are shown in Table 3. It can be seen that the proposed 4D hash grid representation improves the PSNR and SSIM by 6.68 dB and 30.72%, respectively, compared with the six-plane decomposition method. In addition, after adding ST-Former, the PSNR and SSIM are improved by 4.11 dB and 6.78%, respectively. The visualization results can be seen in Figure 3. These results show that our strategy is highly effective.

Sparse view analysis We analyze the reconstruction results of different numbers of projections on the medical dataset to compare the reconstruction performance of different methods under low radiation dose conditions. As shown in Figure 7-a, it can be seen that the proposed method outperforms other methods at any number of projections and is highly robust.

4D Hash grid representation analysis We compared the average quantitative reconstruction results of different hash sizes in the medical dataset, as shown in Figure 7-b. It can be seen that increasing the size of the hash table can improve the reconstruction performance to a certain extent, but the grid features lack constraints and the performance improvement is limited.

ST-Former analysis In the reconstruction results without the ST-Former, we locate the coordinates of the sampling point with and without dot-like artifacts, and input them into the model with the ST-Former to compare their attention maps. As shown in Figure 8, the attention map shows a relatively flat distribution at coordinates without dot-like artifacts. In contrast, at coordinates with dot-like artifacts, the attention map distribution becomes steep. Dot-like artifacts are obvious features of hash conflicts so this experiment can show to a certain extent that ST-Former can guide the model to perform feature selection based on spatiotemporal information, effectively avoiding features with severe hash conflicts.



(a) (b)

Figure 7: (a) Reconstruction results with different numbers of views. (b) Reconstruction results with different hash table sizes, where $log_2 T$ represents the hash table size.

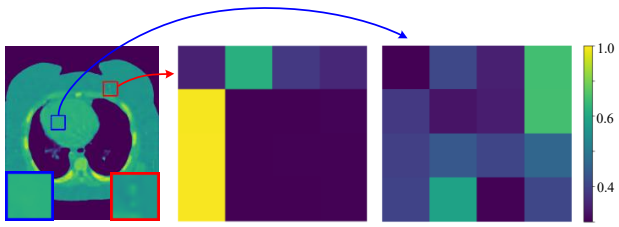


Figure 8: Attention maps of sample points with and without dot-like artifacts. To some extent, it shows that ST-Former guides the model to perform feature selection and avoid dot-like artifacts caused by hash conflicts.

Conclusion

In this paper, we introduce a comprehensive dynamic CT reconstruction model, STNF4D. First, based on the characteristics of dense spatial prediction of dynamic CT reconstruction, we propose the use of 3D hash grids to decompose dynamic scenes. To address the challenges of hash conflicts and the low utilization of grid features, we developed ST-Former, which guides the model in feature selection and optimization by establishing dependencies on spatiotemporal features across different hash grids. We conducted extensive experiments in the medical dataset and industrial dataset. Results show that our method achieves optimal performance in various scenes, significantly outperforming existing state-of-the-art, including learning-based methods.

Acknowledgements

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).

References

Brehm, M.; Paysan, P.; Oelhafen, M.; and Kachelrieß, M. 2013. Artifact-resistant motion estimation with a patient-specific artifact model for motion-compensated cone-beam CT. *Medical physics*, 40(10): 101913.

Cai, Y.; Wang, J.; Yuille, A.; Zhou, Z.; and Wang, A. 2024. Structure-aware sparse-view x-ray 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11174–11183.

Cao, A.; and Johnson, J. 2023. Hexplane: A fast representation for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 130–141.

Chee, G.; O'Connell, D.; Yang, Y.; Singhrao, K.; Low, D.; and Lewis, J. 2019. McSART: an iterative model-based, motion-compensated SART algorithm for CBCT reconstruction. *Physics in Medicine & Biology*, 64(9): 095013.

Chen, S.; Yan, B.; Sang, X.; Chen, D.; Wang, P.; Guo, X.; Zhong, C.; and Wan, H. 2023. Bidirectional optical flow NeRF: high accuracy and high quality under fewer views. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 359–368.

Deng, Z.; Chen, H.; Hu, H.; Xu, Z.; Lyu, T.; Xi, Y.; Chen, Y.; and Zhao, J. 2024. RSTAR: Rotational Streak Artifact Reduction in 4D CBCT using Separable and Circular Convolutions. *arXiv preprint arXiv:2403.16361*.

Deng, Z.; Zhang, W.; Chen, K.; Zhou, Y.; Tian, J.; Quan, G.; and Zhao, J. 2023. 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*.

Fang, J.; Yi, T.; Wang, X.; Xie, L.; Zhang, X.; Liu, W.; Nießner, M.; and Tian, Q. 2022. Fast dynamic radiance fields with time-aware neural voxels. In *SIGGRAPH Asia 2022 Conference Papers*, 1–9.

Feldkamp, L. A.; Davis, L. C.; and Kress, J. W. 1984. Practical cone-beam algorithm. *Josa a*, 1(6): 612–619.

Fridovich-Keil, S.; Meanti, G.; Warburg, F. R.; Recht, B.; and Kanazawa, A. 2023. K-planes: Explicit radiance fields in space, time, and appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12479–12488.

Hugo, G. D.; Weiss, E.; Sleeman, W. C.; Balik, S.; Keall, P. J.; Lu, J.; and Williamson, J. F. 2017. A longitudinal four-dimensional computed tomography and cone beam computed tomography dataset for image-guided radiation therapy research in lung cancer. *Medical physics*, 44(2): 762–771.

Jaffray, D. A.; Siewerdsen, J. H.; Wong, J. W.; and Martinez, A. A. 2002. Flat-panel cone-beam computed tomography for image-guided radiation therapy. *International Journal of Radiation Oncology Biology Physics*, 53(5): 1337–1349.

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

Li, T.; Koong, A.; and Xing, L. 2007. Enhanced 4D cone-beam CT with inter-phase motion model. *Medical physics*, 34(9): 3688–3695.

Liao, Z.; Liu, Y.; Zheng, Q.; and Pan, G. 2024. Spiking NeRF: Representing the Real-World Geometry by a Discontinuous Representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 13790–13798.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.

Mohan, K. A.; Venkatakrishnan, S.; Gibbs, J. W.; Gulsoy, E. B.; Xiao, X.; De Graef, M.; Voorhees, P. W.; and Bouman, C. A. 2015. TIMBIR: A method for time-space reconstruction from interlaced views. *IEEE Transactions on Computational Imaging*, 1(2): 96–111.

Müller, T.; Evans, A.; Schied, C.; and Keller, A. 2022. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (TOG)*, 41(4): 1–15.

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

Pumarola, A.; Corona, E.; Pons-Moll, G.; and Moreno-Noguer, F. 2021. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10318–10327.

Qi, Z.; and Chen, G.-H. 2011. Extraction of tumor motion trajectories using PICCS-4DCBCT: a validation study. *Medical physics*, 38(10): 5530–5538.

Reed, A. W.; Kim, H.; Anirudh, R.; Mohan, K. A.; Champley, K.; Kang, J.; and Jayasuriya, S. 2021. 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*, 2258–2268.

Rückert, D.; Wang, Y.; Li, R.; Idoughi, R.; and Heidrich, W. 2022. Neat: Neural adaptive tomography. *ACM Transactions on Graphics (TOG)*, 41(4): 1–13.

Segars, W. P.; Mahesh, M.; Beck, T. J.; Frey, E. C.; and Tsui, B. M. 2008. Realistic CT simulation using the 4D XCAT phantom. *Medical physics*, 35(8): 3800–3808.

Shao, H.-C.; Mengke, T.; Pan, T.; and Zhang, Y. 2024. Dynamic CBCT imaging using prior model-free spatiotemporal implicit neural representation (PMF-STINR). *Physics in Medicine & Biology*, 69(11): 115030.

Shao, R.; Zheng, Z.; Tu, H.; Liu, B.; Zhang, H.; and Liu, Y. 2023. Tensor4d: Efficient neural 4d decomposition for

high-fidelity dynamic reconstruction and rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16632–16642.

Song, L.; Chen, A.; Li, Z.; Chen, Z.; Chen, L.; Yuan, J.; Xu, Y.; and Geiger, A. 2023. Nerfplayer: A streamable dynamic scene representation with decomposed neural radiance fields. *IEEE Transactions on Visualization and Computer Graphics*, 29(5): 2732–2742.

Vergalasova, I.; and Cai, J. 2020. A modern review of the uncertainties in volumetric imaging of respiratory-induced target motion in lung radiotherapy. *Medical physics*, 47(10): e988–e1008.

Wang, J.; and Gu, X. 2013. Simultaneous motion estimation and image reconstruction (SMEIR) for 4D cone-beam CT. *Medical physics*, 40(10): 101912.

Yang, P.; Ge, X.; Tsui, T.; Liang, X.; Xie, Y.; Hu, Z.; and Niu, T. 2022. Four-dimensional cone beam ct imaging using a single routine scan via deep learning. *IEEE Transactions on Medical Imaging*, 42(5): 1495–1508.

Zeng, R.; Fessler, J. A.; and Balter, J. M. 2007. Estimating 3-D respiratory motion from orbiting views by tomographic image registration. *IEEE Transactions on Medical Imaging*, 26(2): 153–163.

Zha, R.; Zhang, Y.; and Li, H. 2022. NAF: neural attenuation fields for sparse-view CBCT reconstruction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 442–452. Springer.

Zhang, H.; Ma, J.; Bian, Z.; Zeng, D.; Feng, Q.; and Chen, W. 2017. High quality 4D cone-beam CT reconstruction using motion-compensated total variation regularization. *Physics in Medicine & Biology*, 62(8): 3313.

Zhang, Q.; Hu, Y.-C.; Liu, F.; Goodman, K.; Rosenzweig, K. E.; and Mageras, G. S. 2010. Correction of motion artifacts in cone-beam CT using a patient-specific respiratory motion model. *Medical physics*, 37(6Part1): 2901–2909.

Zhang, Y.; Jiang, Z.; Zhang, Y.; and Ren, L. 2024. A review on 4D cone-beam CT (4D-CBCT) in radiation therapy: Technical advances and clinical applications. *Medical Physics*.

Zhang, Y.; Shao, H.-C.; Pan, T.; and Mengke, T. 2023. Dynamic cone-beam CT reconstruction using spatial and temporal implicit neural representation learning (STINR). *Physics in Medicine & Biology*, 68(4): 045005.

Zhi, S.; Kachelrieß, M.; Pan, F.; and Mou, X. 2021. CycN-Net: A convolutional neural network specialized for 4D CBCT images refinement. *IEEE Transactions on Medical Imaging*, 40(11): 3054–3064.