Structure-Aware Sparse-View X-ray 3D Reconstruction

Yuanhao Cai, Jiahao Wang, Zongwei Zhou*, Angtian Wang*, Alan Yuille†
Johns Hopkins University

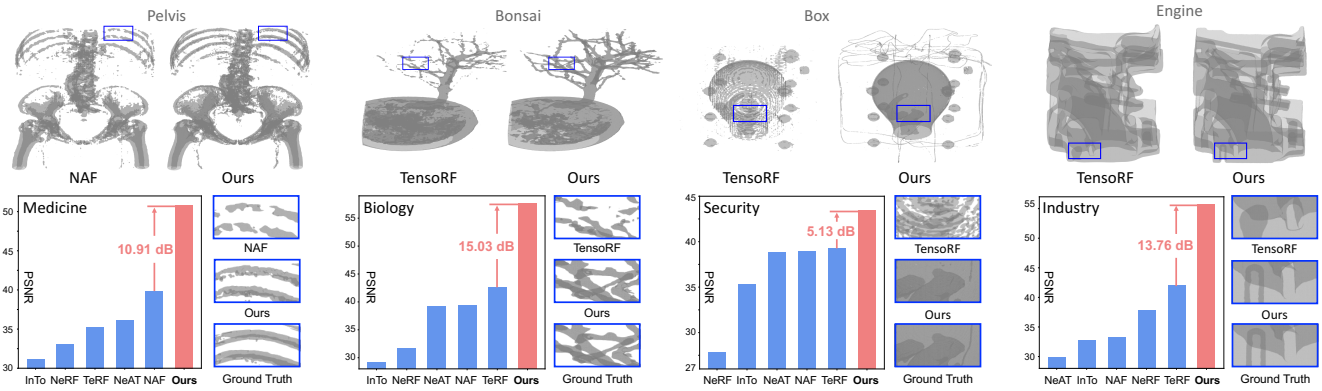


Figure 1. Comparisons of X-ray novel view synthesis. On the collected X3D dataset, our method surpasses state-of-the-art algorithms including InTo (InTomo [56]), NeRF [35], NeAT [43], NAF [59], and TeRF (TensoRF [15]) by **10.91**, **15.03**, **5.13**, and **13.76 dB** in PSNR on the scenes of medicine, biology, security, and industry. The average gains are **over 12 dB**. The visual comparisons of our method and the second-best algorithms on four scenes (pelvis, bonsai, box, and engine) show that our method yields more perceptually pleasing results.

Abstract

*X-ray, known for its ability to reveal internal structures of objects, is expected to provide richer information for 3D reconstruction than visible light. Yet, existing NeRF algorithms overlook this nature of X-ray, leading to their limitations in capturing structural contents of imaged objects. In this paper, we propose a framework, Structure-Aware X-ray Neural Radiodensity Fields (SAX-NeRF), for sparse-view X-ray 3D reconstruction. Firstly, we design a Line Segment-based Transformer (Lineformer) as the backbone of SAX-NeRF. Linefomer captures internal structures of objects in 3D space by modeling the dependencies within each line segment of an X-ray. Secondly, we present a Masked Local-Global (MLG) ray sampling strategy to extract contextual and geometric information in 2D projection. Plus, we collect a larger-scale dataset X3D covering wider X-ray applications. Experiments on X3D show that SAX-NeRF surpasses previous NeRF-based methods by **12.56** and **2.49 dB** on novel view synthesis and CT reconstruction.* https://github.com/caiyuanhao1998/SAX-NeRF

1. Introduction

Compared with natural light, X-ray has stronger penetrating power to reveal more internal structures of imaged objects. Hence, X-ray is widely used for prospective imaging [20, 21, 26, 27] in medicine, biology, security, industry, *etc*. However, X-ray is harmful to human body because of its ionizing radiation. To reduce X-ray exposure, this paper studies the low-dose X-ray 3D reconstruction problem by decreasing X-ray imaging projections in the circular cone beam X-ray scanning scenario [9, 10, 16, 24, 45]. We focus on two tasks, *i.e.*, novel view synthesis (NVS) and computed tomography (CT) reconstruction. NVS aims to create new projections of a scene from viewpoints not originally captured. CT reconstruction retrieves the 3D CT volume of the scanned object from multi-view X-ray projections. These two tasks are complementary with an overall objective to reconstruct 3D representations from 2D projections.

A majority of existing deep learning-based methods employ a powerful model such as convolutional neural network (CNN) to learn a brute-force mapping from 2D X-ray projections to 3D CT volumes. These methods require a large number of projection-CT pairs for training. Yet, CT volumes are not accessible in practice. Collecting even a small projection-CT dataset is tedious, labor-intensive, and harmful to health. Plus, these paired learning-based methods fail in generalizing from one application to another due to the large domain discrepancy between different CT datasets.

Recently, the emergence of NeRF [35] provides a more

* = corresponding authors. † = project leader.

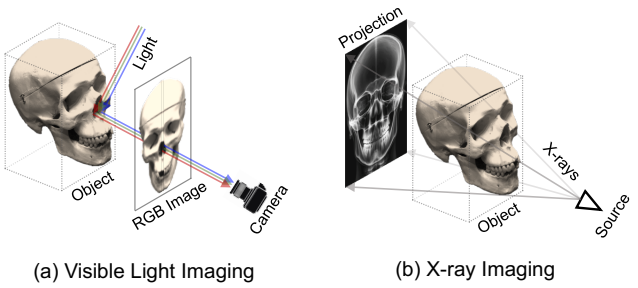


Figure 2. Visible light *vs.* X-ray. Visible light imaging relies on reflection. X-ray imaging is based on penetration and attenuation.

reasonable solution to X-ray 3D reconstruction. Compared with paired learning-based algorithms, NeRF-based methods do not require CT volumes for training. Instead, they only need projections of just one scene. Although RGB NeRF algorithms have been well developed, directly applying them for X-ray scenes may achieve suboptimal results due to the fundamental differences between visible light and X-ray imaging. As compared in Fig. 2, visible light imaging relies on the reflection off the surface of an object. It mainly captures external features. In contrast, X-rays penetrate the object and attenuate, thereby forming an image. X-ray imaging primarily reveals internal structures, which provide key clues for X-ray 3D reconstruction.

Nonetheless, current NeRF-based methods overlook this critical property of X-ray imaging. **Firstly**, they learn NeRF by a simple multilayer perceptron (MLP). X-ray attenuates differently when penetrating different structures. However, MLP treats each point on an X-ray equally, showing limitations in modeling 3D structures of objects. **Secondly**, previous methods mainly adopt a naive pixel-level ray sampling strategy in the training phase. They randomly sample X-rays corresponding to scattered pixels on the whole image coordinate system. As a result, the contextual information and geometric structures in 2D projection are not well extracted. Plus, X-ray projections are spatially sparse. Sampling X-rays on uninformative regions may lead to low efficiency. **Besides**, existing methods mainly study X-ray 3D reconstruction in limited medical scenes while their performance on other applications is still under-explored.

To tackle these issues, we propose a novel framework, Structure-Aware X-ray Neural Radiodensity Fields (SAX-NeRF), with the key insight of capturing 2D and 3D structures in X-ray imaging. **Firstly**, we design a Line Segment-based Transformer (Lineformer) as the backbone of SAX-NeRF. It partitions an X-ray into different line segments and then samples points on each one. By computing self-attention within every piece of the X-ray, Lineformer can model internal dependencies and learn complex 3D structures of different parts penetrated by the X-ray. Unlike vanilla Transformer [48] whose computational cost is quadratic to the number of input points, Lineformer is more efficient by enjoying linear computational complexity. **Secondly**, we present a Masked Local-Global (MLG) ray sampling strategy. It uses a binary mask to segment informative foreground regions on the projection. We crop non-overlapping patches from these informative regions and then sample X-rays that land on the pixels inside these patches to help Lineformer perceive local contextual information and 2D structures. For the informative regions outside the patches, we randomly sample X-rays to help Lineformer perceive the scene's 2D global shape and geometry. **Besides**, we collect a larger-scale dataset, X3D, to evaluate the performance of X-ray 3D reconstruction algorithms in wider application scenarios. As shown in Fig. 1, our SAX-NeRF surpasses state-of-the-art (SOTA) NeRF-based methods by large margins on the NVS task. The average improvements on all scenes of X3D are **over 12 dB**.

Our contributions can be summarized as follows:

- We propose a novel method, SAX-NeRF, for sparse-view X-ray 3D reconstruction without CT data for training.
- We present a new Transformer, Lineformer, to capture complex internal structures of imaged objects in 3D space. To our knowledge, it is the first attempt to explore the potential of Transformer in X-ray neural rendering.
- We design an MLG sampling strategy to extract geometric and contextual information of objects in 2D projection.
- We establish a larger-scale benchmark, X3D, for X-ray 3D reconstruction. Experiments show that our method outperforms SOTA methods on NVS and CT reconstruction tasks across different application scenarios of X-ray.

2. Related Work

2.1. Neural Rendering

NeRF [35] represents objects via an implicit function of color and volume density, yielding high-quality results on the NVS task. Follow-up works improve NeRF with more fine-grained details [6, 7] and broader applications [17, 49, 52]. Meanwhile, to reduce the computational cost of NeRF, learnable feature encodings [15, 18, 36] are designed to embed input point positions. However, applying existing RGB NeRF methods for X-ray rendering [22, 43, 59] may achieve suboptimal results due to the differences between visible light and X-ray imaging. For instance, NAF [59] follows NeRF to employ an MLP model for medical X-ray neural rendering, showing limitations in capturing complex structures of imaged objects in 3D space and 2D projection.

2.2. Cone Beam CT Reconstruction

Traditional cone beam CT Reconstruction algorithms are mainly divided into two categories: analytical methods [25, 54] and optimization-based methods [2, 34, 37, 44, 46, 57]. Analytical methods predict the CT volume by solving the Radon transformation [40] and its inverse. These methods

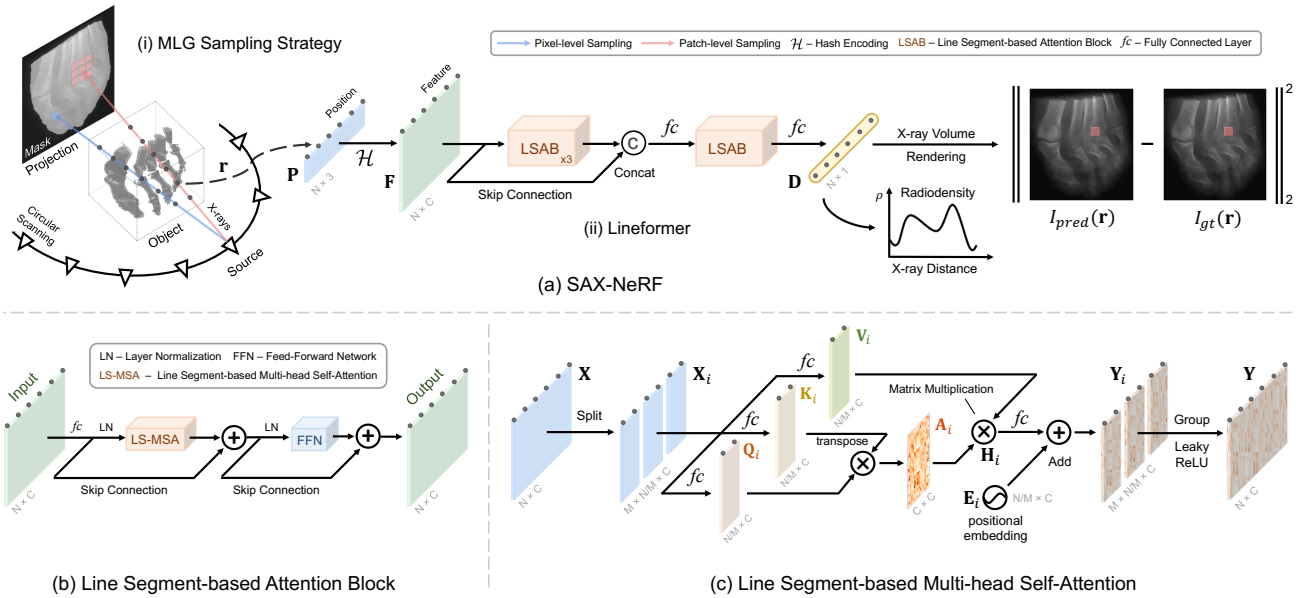


Figure 3. Overview of our method. (a) SAX-NeRF uses (i) MLG strategy to sample an X-ray batch $\mathcal{R}$. Then N point positions $\mathbf{P}$ on each X-ray $\mathbf{r} \in \mathcal{R}$ are sampled and input into (ii) Lineformer to produce the radiodensity $\mathbf{D}$. (b) Line Segment-based Attention Block (LSAB) is the basic unit of Lineformer. It captures inner structural dependencies by (c) Line Segment-based Multi-head Self-Attention (LS-MSA).

can achieve good results when given hundreds of projections but fail in handling sparse-view cases. Optimization-based algorithms treat the reconstruction as a *maximum a posteriori* (MAP) problem based on hand-crafted image priors and solve it by iteratively minimizing the energy function, which takes a long time. Recently, CNNs [3, 30, 33, 53, 56] and diffusion models [19] have been applied to CT reconstruction and achieve good results. Yet, these methods require a number of data pairs for training. To avoid the above restrictions, we develop NeRF-based algorithms.

2.3. Vision Transformer

Transformer [48] is first proposed for machine translation. In recent years, it has achieved great success in computer vision including image classification [1, 5, 23], object detection [47, 58, 63], semantic segmentation [51, 61, 62], image restoration [12–14, 55] and generation [28, 29, 60], *etc.* Nonetheless, directly applying vanilla Transformer for X-ray neural rendering will suffer from expensive computational cost with respect to the number of input points. The potential of Transformer for X-ray neural rendering still remains under-explored. We aim to fill this research gap.

3. Method

3.1. Overall Framework

Fig. 3 illustrates the pipeline of our method. The left part of Fig. 3 (a) depicts the scenario of circular cone beam X-ray scanning where a scanner emits cone-shaped X-ray beams and captures sparse-view projections at equal angular intervals. We first use (i) Masked Local-Global (MLG) strategy to sample a batch of X-rays $\mathcal{R}$ landing on the projections for training. Then N point positions $\mathbf{P} = \{\mathbf{p}_1, \cdots, \mathbf{p}_N\} \in \mathbb{R}^{N\times 3}$ are sampled on each X-ray $\mathbf{r} \in \mathcal{R}$ and fed into (ii) Lineformer. The basic unit of Lineformer is Line Segment-based Attention Block (LSAB). As shown in Fig. 3 (b), an LSAB consists of a fully connected (fc) layer, two layer normalization (LN), a feed-forward network (FFN), and a Line Segment-based Multi-head Self-Attention (LS-MSA). The details of LS-MSA are depicted in Fig. 3 (c).

Firstly, we review RGB NeRF. An MLP with weights Θ is usually employed to learn the mapping function F_Θ from the point position $(x, y, z) \in \mathbb{R}^3$ at the view direction (θ, ϕ) to the color $(R, G, B) \in \mathbb{R}^3$ and volume density $\sigma \in \mathbb{R}$ as

$$F_\Theta : (x, y, z, \theta, \phi) \rightarrow (R, G, B, \sigma). \tag{1}$$

As shown in Fig. 2, visible light of specific wavelengths reflects off the surface of the object, thus revealing its color. In contrast, X-rays penetrate the object, thereby not reflecting the color information. Instead, it records the radiodensity property that denotes the degree to which a substance blocks or attenuates the passage of X-rays or other ionizing radiation. Since the radiodensity only depends on the point position, we aim to model the neural radiodensity fields as

$$F_{\Theta_L} : (x, y, z) \rightarrow \rho, \tag{2}$$

where F_{Θ_L} represents the mapping function of our Lineformer with weights Θ_L and $\rho \in \mathbb{R}$ denotes the radiodensity. According to the Beer-Lambert law, the intensity of an X-ray is reduced by the exponential integration of the

traversed object's radiodensity. Hence, the ground-truth intensity $I_{gt}(\mathbf{r}) \in \mathbb{R}$ of the X-ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d} \in \mathbb{R}^3$ with the near and far bounds t_n and $t_f \in \mathbb{R}$ can be formulated as

$$I_{gt}(\mathbf{r}) = I_0 \cdot \exp\big(-\int_{t_n}^{t_f} \rho(\mathbf{r}(t))dt\big), \tag{3}$$

where I_0 is the initial intensity. By discretizing Eq. (3), we derive the predicted projection intensity $I_{pred}(\mathbf{r}) \in \mathbb{R}$ as

$$I_{pred}(\mathbf{r}) = I_0 \cdot \exp\big(-\sum_{i=1}^{N} \rho_i \delta_i\big), \tag{4}$$

where ρ_i denotes the predicted radiodensity of the i-th sampled point and $\delta_i = ||\mathbf{p}_{i+1} - \mathbf{p}_i||$ is the distance between adjacent points. Eventually, the training objective is to minimize the total squared error $\mathcal{L}$ between the predicted and ground-truth intensities in the training X-ray batch $\mathcal{R}$ as

$$\mathcal{L} = \sum_{\mathbf{r}\in\mathcal{R}} \Big|\Big| I_{pred}(\mathbf{r}) - I_{gt}(\mathbf{r}) \Big|\Big|_2^2, \tag{5}$$

where $I_{gt}(\mathbf{r})$ is obtained from the pixel value on projection. $\mathcal{L}$ is depicted by the red pixels in the right part of Fig. 3 (a).

3.2. Line Segment-based Transfomer

As aforementioned, X-ray imaging reveals internal structures of imaged objects, which provide key clues for 3D reconstruction. Yet, previous methods overlook this important imaging property. Specifically, similar to RGB NeRF algorithms, existing X-ray NeRF methods [56, 59] mainly adopt a simple MLP model to learn the implicit neural representations. X-ray attenuates differently when penetrating different structural contents. Yet, the MLP model treats each sampled point on an X-ray equally, showing limitations in modeling the 3D structures penetrated by the X-ray.

Towards this issue, we propose a Line Segment-based Transformer (Lineformer), as shown in Fig. 3 (a) (ii). The point position $\mathbf{P}$ is firstly fed into a hash encoding [36] module $\mathcal{H}$ to produce point feature $\mathbf{F} \in \mathbb{R}^{N\times C}$ as $\mathbf{F} = \mathcal{H}(\mathbf{P})$. Then $\mathbf{F}$ undergoes four LSABs with a skip connection and two fc layers to derive the point radiodensity $\mathbf{D} \in \mathbb{R}^N$.

LSAB is the basic unit of Lineformer. Its most important component is the LS-MSA mechanism, which captures internal structural dependencies by computing self-attention within each line segment of an X-ray. As illustrated in Fig. 3 (c), the input point feature $\mathbf{X} \in \mathbb{R}^{N\times C}$ is firstly partitioned into M segments along the point dimension as

$$\mathbf{X} = [\mathbf{X}_1,\ \mathbf{X}_2,\ \cdots,\ \mathbf{X}_M]^{\mathrm{T}}, \tag{6}$$

where $\mathbf{X}_i \in \mathbb{R}^{\frac{N}{M}\times C}$ and $i = 1, 2, \cdots, M$. Then each $\mathbf{X}_i$ is linearly projected into *query* $\mathbf{Q}_i \in \mathbb{R}^{\frac{N}{M}\times C}$, *key* $\mathbf{K}_i \in \mathbb{R}^{\frac{N}{M}\times C}$, and *value* $\mathbf{V}_i \in \mathbb{R}^{\frac{N}{M}\times C}$ by three fc layers as

$$\mathbf{Q}_i = \mathbf{X}_i\mathbf{W}^{\mathbf{Q}_i},\quad \mathbf{K}_i = \mathbf{X}_i\mathbf{W}^{\mathbf{K}_i},\quad \mathbf{V}_i = \mathbf{X}_i\mathbf{W}^{\mathbf{V}_i}, \tag{7}$$

where $\mathbf{W}^{\mathbf{Q}_i}$, $\mathbf{W}^{\mathbf{K}_i}$, and $\mathbf{W}^{\mathbf{V}_i} \in \mathbb{R}^{C\times C}$ are learnable parameters of the fc layers; $biases$ are omitted for simplification. Subsequently, $\mathbf{W}^{\mathbf{Q}_i}$, $\mathbf{W}^{\mathbf{K}_i}$, and $\mathbf{W}^{\mathbf{V}_i}$ are uniformly split into k heads along the channel dimension as

$$\begin{aligned}\mathbf{Q}_i &= [\mathbf{Q}_i^1,\ \mathbf{Q}_i^2,\ \cdots,\ \mathbf{Q}_i^k],\\ \mathbf{K}_i &= [\mathbf{K}_i^1,\ \mathbf{K}_i^2,\ \cdots,\ \mathbf{K}_i^k],\\ \mathbf{V}_i &= [\mathbf{V}_i^1,\ \mathbf{V}_i^2,\ \cdots,\ \mathbf{V}_i^k].\end{aligned} \tag{8}$$

The dimension for each head is $d_h = C/k$. Fig. 3 (b) illustrates the situation with $k = 1$ for simplicity. Then the self-attention within each head $\mathbf{H}_i^j$ is computed as

$$\mathbf{H}_i^j = \mathrm{Attn}(\mathbf{Q}_i^j, \mathbf{K}_i^j, \mathbf{V}_i^j) = \mathbf{V}_i^j\ \mathrm{softmax}(\frac{{\mathbf{K}_i^j}^{\mathrm{T}}\mathbf{Q}_i^j}{\alpha_i^j}), \tag{9}$$

where $\alpha_i^j \in \mathbb{R}$ is a learnable parameter that adaptively scales the inner product before the softmax function. Successively, k heads are concatenated in channel dimension to pass through an fc layer and then plus a positional embedding $\mathbf{E}_i \in \mathbb{R}^{\frac{N}{M}\times C}$ to derive the i-th output $\mathbf{Y}_i \in \mathbb{R}^{\frac{N}{M}\times C}$ as

$$\mathbf{Y}_i = [\mathbf{H}_i^1,\ \mathbf{H}_i^2,\ \cdots,\ \mathbf{H}_i^k]\ \mathbf{W}_i + \mathbf{E}_i, \tag{10}$$

where $\mathbf{W}_i \in \mathbb{R}^{C\times C}$ are learnable parameters of the fc layer. Finally, we group the outputs of M segments in point dimension to obtain the output feature $\mathbf{Y} \in \mathbb{R}^{N\times C}$ as

$$\mathbf{Y} = [\mathbf{Y}_1,\ \mathbf{Y}_2,\ \cdots,\ \mathbf{Y}_M]^{\mathrm{T}}. \tag{11}$$

By capturing the interactions of points within each line segment, the proposed Lineformer is more capable of perceiving the complex internal 3D structures of different parts penetrated by the X-ray and therefore modeling the implicit neural radiodensity fields in Eq. (2) more accurately.

Complexity Analysis. We analyze the computational complexity of our LS-MSA and compare it with the global multi-head self-attention (G-MSA) mechanism of vanilla Transformer. The computational cost of LS-MSA primarily comes from the two matrix multiplication, *i.e.*, $\mathbb{R}^{d_h\times\frac{N}{M}} \times \mathbb{R}^{\frac{N}{M}\times d_h}$ and $\mathbb{R}^{\frac{N}{M}\times d_h} \times \mathbb{R}^{d_h\times d_h}$, in Eq. (9) performed $k \times M$ times. Thus, the complexity of LS-MSA is formulated as

$$\begin{aligned}\mathcal{O}(\text{LS-MSA}) &= kM \cdot [d_h \cdot (d_h \cdot \frac{N}{M}) + \frac{N}{M} \cdot (d_h \cdot d_h)],\\ &= 2kMd_h^2\frac{N}{M} = 2Nk(\frac{C}{k})^2 = \frac{2NC^2}{k}.\end{aligned} \tag{12}$$

While the complexity of G-MSA is formulated as

$$\mathcal{O}(\text{G-MSA}) = 2N^2C. \tag{13}$$

Compare Eq. (12) with Eq. (13). $\mathcal{O}(\text{G-MSA})$ is quadratic to the number of input points (N). This heavy computational burden impedes the application of Transformer for X-ray 3D reconstruction. In contrast, $\mathcal{O}(\text{LS-MSA})$ is linear to N. This significantly reduced cost allows for the integration of LS-MSA into each basic unit LSAB of Lineformer, thereby further exploring the tremendous potential of Transformer.

3.3. Masked Local-Global Ray Sampling

As shown in Fig. 4 (a), existing NeRF algorithms mainly adopt a naive pixel-level ray sampling strategy. They randomly sample X-rays corresponding to scattered pixels on the whole image coordinate system for training. This naive strategy has two drawbacks. **Firstly**, it shows limitations in extracting local contextual and geometric representations in 2D projection because the semantic information from neighbor pixels is not captured. **Secondly**, X-ray images are spatially sparse. Some randomly sampled X-rays may land on the background dark regions of the projection, such as the pixel p_{bg} in Fig. 4 (a). These X-rays do not penetrate the object and thus are not imaged on the projection. In other words, these X-rays are uninformative because they do not characterize the radiodensity property of the object. Learning with these X-rays will degrade the model efficiency.

To address these problems, we propose a Masked Local-Global (MLG) ray sampling strategy, as shown in Fig. 4 (b). MLG first uses a mask $\mathbf{M} \in \mathbb{R}^{H \times W}$ to segment the imaged foreground regions. $\mathbf{M}$ is derived by binarizing the projection $\mathbf{I} \in \mathbb{R}^{H \times W}$ with a threshold $T \in \mathbb{R}$ as $\mathbf{M} = \mathbf{1}_{\mathbf{I}>T}$. Subsequently, to avoid redundant sampling, we partition $\mathbf{M}$ into a set $\mathcal{W} \in \mathbb{R}^{\frac{HW}{S^2} \times S \times S}$ of $\frac{HW}{S^2}$ non-overlapping windows with size $S \times S$. Let $\mathcal{W}_f$ denote the set of windows that are entirely contained in the foreground regions as

$$\mathcal{W}_f = \{\mathbf{w} \in \mathcal{W} \mid \mathbf{w} = \mathbf{1}_{S \times S}\}. \tag{14}$$

To capture local semantic information of the object, we perform patch-level sampling. Specifically, we randomly select N_l windows $\mathcal{W}_l = \{\mathbf{w}_1, \cdots, \mathbf{w}_{N_l}\}$ from $\mathcal{W}_f$, as shown in the red patches of Fig. 4 (b). Then the X-ray set $\mathcal{R}_l$ corresponding to the pixels within $\mathcal{W}_l$ can be formulated as

$$\mathcal{R}_l = \bigcup_{i=1}^{N_l} \bigcup_{p \in \mathbf{w}_i} \text{Ray}(p), \tag{15}$$

where $\text{Ray}(p)$ is a function that maps from a pixel p to its corresponding X-ray. Furthermore, to assist the model in better capturing global contextual representations and perceiving the overall geometric shape of the imaged object, we perform pixel-level sampling. Particularly, we randomly select N_g pixels $\mathcal{P}$ from the foreground regions excluding the area of $\mathcal{W}_l$ to avoid repeated ray sampling, as depicted in the blue pixels of Fig. 4 (b). $\mathcal{P}$ can be formulated as

$$\mathcal{P} = \{p \in (\mathbf{M} - \mathcal{W}_l) \mid p = 1\}. \tag{16}$$

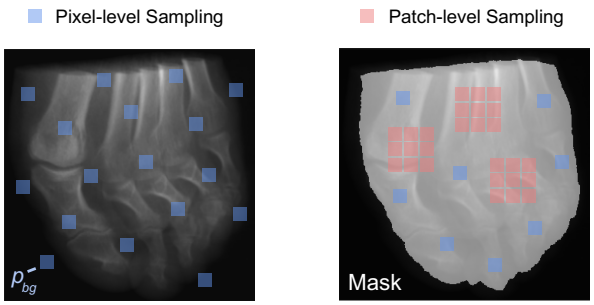
Then the X-ray set $\mathcal{R}_g$ corresponding to $\mathcal{P}$ is obtained by

$$\mathcal{R}_g = \bigcup_{p \in \mathcal{P}} \text{Ray}(p). \tag{17}$$

Finally, the training X-ray batch is the union of $\mathcal{R}_l$ and $\mathcal{R}_g$:

$$\mathcal{R} = \mathcal{R}_l \bigcup \mathcal{R}_g. \tag{18}$$

Using MLG ray sampling strategy, the model can more effectively capture the contextual information and model the geometric structures of the imaged object on 2D projection.



(a) Naive Sampling (b) Our MLG Sampling

Figure 4. Comparison of ray sampling. (a) The naive strategy samples X-rays that land on scattered pixels. (b) Our MLG strategy performs pixel- and patch-level sampling on foreground regions.

4. Experiment

4.1. Experimental Settings

X3D Dataset. Previous methods mainly conduct X-ray 3D reconstruction research on limited medical applications. For example, NAF [59] is evaluated on five medical scenes. The performance of NeRF-based methods on other X-ray applications is under-explored. To fill this research gap, we collect a larger-scale dataset, X3D, containing 15 scenes and covering 4 applications, *i.e.*, medicine, biology, security, and industry. We collect the CT volumes of X3D from public datasets. Specifically, the chest, backpack, carp, and pancreas datasets are collected from LIDC-IDRI [4], MIDA [11], D^{2}VR [41], and DeepOrgan [42], respectively. The teapot, aneurism, bonsai, and foot datasets are obtained from VOLVIS [39] and the rest are from the open scientific visualization dataset [32]. Then we use the tomographic method TIGRE [8] to generate projections by scanning CT volumes with 3% noise in the range of $0° \sim 180°$.

Implementation Details. We implement our SAX-NeRF by PyTorch [38]. The model is trained with the Adam [31] optimizer ($\beta_1 = 0.9$ and $\beta_2 = 0.999$) for 3000 iterations. The learning rate is initially set to 1×10^{-4} and is halved every 1500 iterations during the training procedure. The batch size of X-rays is set to 2048, 1024 of which is from patch-level sampling and the other 1024 is from pixel-level sampling. We uniformly sample 320 points along each X-ray. For each scene, we use its 50 projections to train, another 50 projections to test the performance of NVS, and its CT volume to evaluate the results of CT reconstruction. All experiments are conducted on an RTX 8000 GPU. We adopt the peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) [50] as the evaluation metrics.

4.2. Main Results

Novel View Synthesis. Tab. 1 lists the quantitative results of PSNR and SSIM on the NVS task. We compare our

Method	InTomo [56]		NeRF [35]		NeAT [43]		TensoRF [15]		NAF [59]		**SAX-NeRF**	
Scene	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Jaw	26.91	0.9937	34.84	0.9981	32.68	0.9911	34.06	0.9964	39.89	0.9988	**42.75**	**0.9992**
Leg	42.53	0.9976	45.92	0.9989	47.71	0.9981	41.40	0.9969	50.87	0.9988	**56.86**	**0.9996**
Box	34.65	0.9963	35.67	0.9985	36.14	0.9957	35.43	0.9977	35.98	0.9955	**39.67**	**0.9992**
Carp	24.04	0.9648	20.62	0.9467	31.26	0.9620	37.35	0.9973	29.60	0.9593	**59.88**	**0.9999**
Foot	39.48	0.9979	41.05	0.9989	38.24	0.9963	37.73	0.9929	38.35	0.9913	**46.64**	**0.9994**
Head	34.83	0.9977	29.76	0.9991	27.74	0.9295	34.43	0.9878	30.17	0.9531	**53.06**	**0.9995**
Pelvis	38.72	0.9961	40.79	0.9972	37.70	0.9866	41.57	0.9948	43.76	0.9975	**53.27**	**0.9995**
Chest	28.95	0.9915	36.16	0.9988	40.77	0.9990	23.61	0.9402	42.37	0.9993	**47.42**	**0.9994**
Bonsai	39.26	0.9953	37.67	0.9983	47.02	0.9985	47.80	0.9989	49.03	0.9989	**55.33**	**0.9995**
Teapot	41.51	0.9978	34.66	0.9993	29.29	0.9669	44.18	0.9993	34.92	0.9985	**52.62**	**0.9996**
Engine	23.99	0.9517	21.07	0.9334	30.36	0.8854	39.72	0.9918	31.68	0.9195	**58.80**	**0.9998**
Pancreas	20.03	0.8537	19.85	0.8560	37.53	0.9017	29.24	0.8031	36.23	0.8844	**49.88**	**0.9978**
Abdomen	27.64	0.9646	24.62	0.9559	26.74	0.8563	27.38	0.8730	37.59	0.9855	**54.22**	**0.9996**
Aneurism	20.81	0.9621	24.97	0.9792	35.41	0.9936	47.99	0.9997	39.62	0.9990	**52.91**	**0.9998**
Backpack	36.09	0.9918	39.75	0.9962	41.60	0.9969	43.16	0.9977	42.02	0.9982	**47.17**	**0.9989**
Average	31.96	0.9768	32.49	0.9770	36.01	0.9638	37.67	0.9712	38.81	0.9785	**51.37**	**0.9994**

Table 1. Quantitative comparisons on the novel view synthesis task. The best results are in **bold** and the second-best results are underlined.

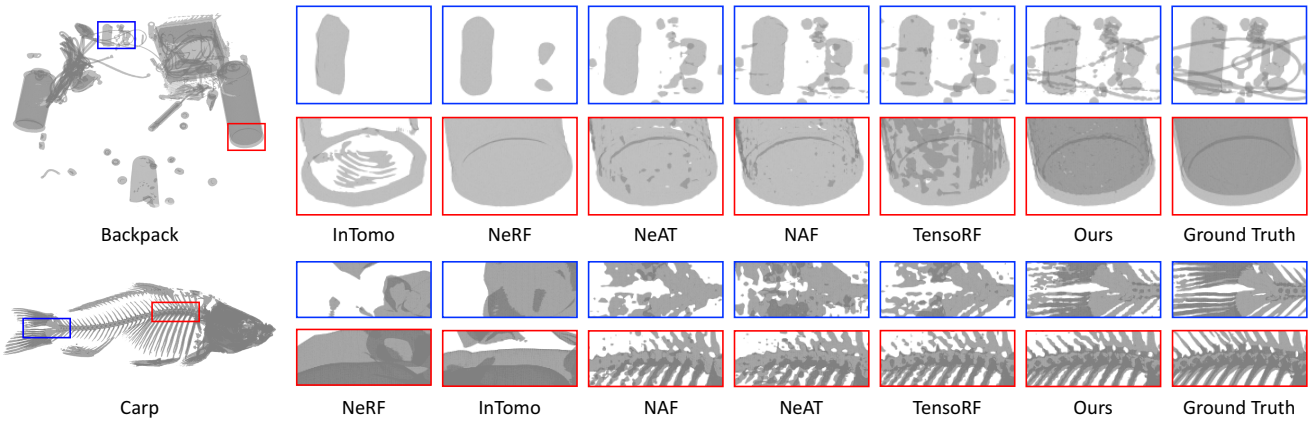


Figure 5. Qualitative results of novel view synthesis on the scenes of backpack (top) and carp (bottom). Please zoom in for a better view.

SAX-NeRF with five SOTA NeRF-based algorithms including InTomo [56], NeRF [35], NeAT [43], TensoRF [15], and NAF [59]. The input and output of all methods are set the same as Eq. (2) for fair comparison. It can be observed that our SAX-NeRF significantly outperforms SOTA methods on all scenes. Specifically, when compared with the recent best general RGB NeRF algorithm TensoRF, our SAX-NeRF is 13.70 dB (51.37 - 37.67) and 0.0282 (0.9994 - 0.9712) higher in PSNR and SSIM. When compared with the recent best medical NeRF method NAF, SAX-NeRF surpasses it by 12.56 dB in PSNR and 0.0209 in SSIM. The average improvements of our method on the scenes of medicine, biology, security, and industry are 10.91, 15.03, 5.13, and 13.76 dB, as shown in the bar charts of Fig. 1.

The qualitative results are depicted in Fig. 5. As can be observed from the zoomed-in patches, previous methods are less effective in synthesizing novel projections. They either produce blurry images or fail to reconstruct structural contents. In contrast, our SAX-NeRF yields more visually pleasing results with clearer textures and more fine-grained details while preserving more complete geometric structures. More visual comparisons are shown in Fig. 1.

CT Reconstruction. Tab. 2 reports the quantitative results on the CT reconstruction task. For fairness, we do not compare projection-CT paired learning-based algorithms, but instead focus on comparing methods that only require X-ray projections of single scenes for training or direct processing. In addition to the five SOTA NeRF-based algorithms. We also compare SAX-NeRF with an analytical method (FDK [25]) and two optimization-based algorithms (ASD-POCS [46] and SART [2]). Our method yields the best results on all scenes. In particular, SAX-NeRF dramatically outperforms previous NeRF-based, optimization-based, and analytical algorithms by over 2.49, 4.92, and 12.13 dB.

Fig. 6 displays the visual comparisons in four application scenarios including medicine (head), biology (carp), security (box), and industry (teapot). Other methods either produce over-smooth images blurring the structural contents or

Method	FDK [25]		ASD-POCS [46]		SART [2]		InTomo [56]		NeRF [35]		NeAT [43]		TensoRF [15]		NAF [59]		**SAX-NeRF**	
Scene	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Jaw	28.58	0.7816	33.25	0.9325	33.13	0.9301	31.95	0.9162	32.17	0.9114	32.53	0.9139	31.90	0.8971	34.14	0.9358	**35.47**	**0.9525**
Leg	28.48	0.6690	35.39	0.9826	35.30	0.9809	36.41	0.9882	39.27	0.9938	40.29	0.9902	40.70	0.9923	41.28	0.9940	**43.47**	**0.9973**
Box	24.14	0.5616	31.27	0.9226	31.20	0.9200	30.59	0.9140	33.58	0.9494	31.58	0.9298	32.17	0.9314	31.78	0.9309	**35.33**	**0.9602**
Carp	32.32	0.8177	37.63	0.9777	36.89	0.9682	32.47	0.9493	32.99	0.9529	36.85	0.9576	37.52	0.9687	37.93	0.9711	**42.72**	**0.9902**
Foot	24.53	0.6000	29.98	0.9208	30.29	0.9296	31.43	0.9127	30.03	0.9072	30.86	0.9221	30.46	0.9153	31.63	0.9363	**32.25**	**0.9403**
Head	26.17	0.7155	35.27	0.9707	34.88	0.9597	31.07	0.9303	34.15	0.9672	35.56	0.9679	35.53	0.9672	36.46	0.9743	**39.70**	**0.9888**
Pelvis	26.91	0.6367	34.26	0.9493	34.38	0.9481	30.38	0.9042	31.72	0.9170	33.73	0.9370	35.13	0.9528	36.01	0.9654	**40.40**	**0.9870**
Chest	22.89	0.7861	31.13	0.9422	32.17	0.9594	22.04	0.7460	28.40	0.8925	31.20	0.9497	30.13	0.9308	33.05	0.9581	**34.38**	**0.9718**
Bonsai	24.53	0.7276	32.70	0.9529	33.02	0.9600	28.90	0.8811	31.77	0.9382	33.20	0.9476	33.47	0.9521	33.85	0.9585	**36.51**	**0.9761**
Teapot	31.07	0.8059	37.35	0.9800	37.38	0.9787	36.15	0.9786	41.67	0.9945	40.85	0.9872	42.71	0.9942	42.56	0.9926	**44.32**	**0.9970**
Engine	23.02	0.5405	30.81	0.9580	30.44	0.9442	27.49	0.9264	36.85	0.9858	36.63	0.9804	35.21	0.9728	37.84	0.9859	**38.77**	**0.9917**
Pancreas	9.641	0.1232	18.30	0.7701	18.36	0.7008	16.01	0.7865	17.73	0.8614	19.06	0.8541	19.75	0.7737	19.41	0.8126	**22.98**	**0.9531**
Abdomen	22.63	0.6030	31.46	0.9231	31.40	0.9170	28.05	0.8754	29.71	0.9049	31.14	0.9060	31.51	0.9073	34.45	0.9501	**35.01**	**0.9598**
Aneurism	28.07	0.7295	34.73	0.9864	34.76	0.9864	30.32	0.9652	31.97	0.9353	35.80	0.9819	37.36	0.9889	37.73	0.9871	**41.46**	**0.9956**
Backpack	23.84	0.5351	31.34	0.9309	31.32	0.9294	28.77	0.8753	30.28	0.9192	31.90	0.9345	33.16	0.9362	33.26	0.9501	**35.97**	**0.9688**
Average	25.12	0.6422	32.32	0.9400	32.33	0.9342	30.29	0.9189	32.15	0.9354	33.41	0.9447	33.78	0.9387	34.76	0.9535	**37.25**	**0.9753**

Table 2. Quantitative comparisons on the CT reconstruction task. The best results are in **bold** and the second-best results are underlined.

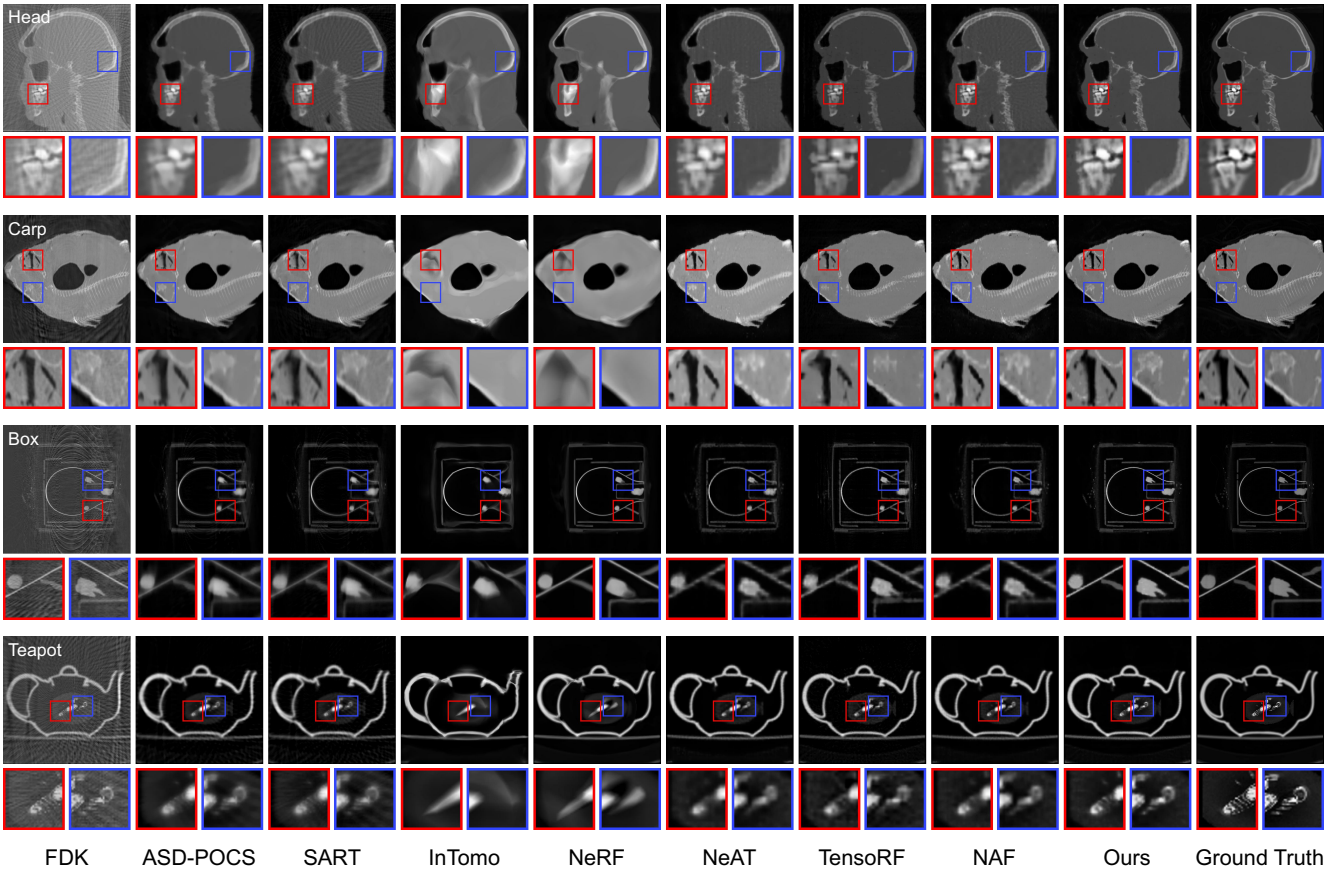


Figure 6. Visual results of CT reconstruction on the scenes of head, carp, box, and teapot (top to bottom). Please zoom in for a better view.

introduce distracting artifacts. In contrast, our SAX-NeRF is more favorable to reconstruct vivid high-frequency details such as sharp edges while maintaining spatial smoothness of homogeneous regions within complex structures.

These results convincingly demonstrate the advantages of the proposed SAX-NeRF in X-ray 3D reconstruction.

4.3. Ablation Study

To reliably evaluate the effectiveness of our approaches, we conduct ablation study on all scenes of X3D and report the average PSNR / SSIM results in the following part.

Break-down Ablation. We adopt a baseline model that

Baseline	LS-MSA	MLG	NVS	CT
✓			37.97 / 0.9748	34.21 / 0.9513
✓	✓		47.95 / 0.9945	36.86 / 0.9717
✓		✓	43.51 / 0.9861	35.30 / 0.9601
✓	✓	✓	**51.37 / 0.9994**	**37.25 / 0.9753**

(a) Break-down ablation to higher performance.

Num	20	40	80	160	320
NVS	45.776	47.832	50.148	**51.365**	50.620
	0.9892	0.9939	0.9991	**0.9994**	0.9992
CT	35.845	36.787	37.088	**37.249**	37.186
	0.9670	0.9711	0.9745	**0.9753**	0.9749

(b) Analysis of the line segment quantity.

Size	2×2	4×4	8×8	16×16	32×32
NVS	48.515	**51.365**	50.297	50.408	49.629
	0.9976	**0.9994**	0.9992	0.9992	0.9989
CT	36.982	**37.249**	37.118	37.163	37.014
	0.9733	**0.9753**	0.9748	0.9749	0.9741

(c) Analysis of the sampling patch size.

Table 3. We conduct ablation study on all scenes of X3D. Average PSNR and SSIM are reported on the NVS and CT reconstruction tasks.

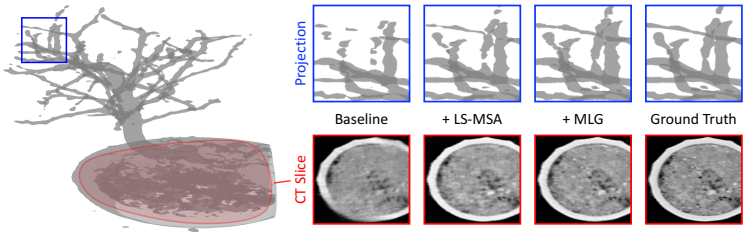


Figure 7. Visual analysis. Using LS-MSA and MLG captures more structures.

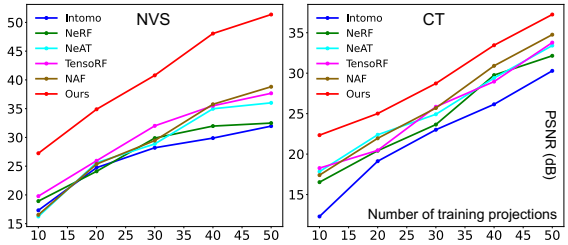


Figure 8. Analysis of the number of training projections.

is derived by directly removing the LS-MSA module and MLG sampling strategy from our SAX-NeRF to conduct the break-down ablation study. The results are listed in Tab. 3a. The baseline model yields 37.97 and 34.21 dB on NVS and CT reconstruction. When using LS-MSA, the baseline model gains by 9.98 and 2.65 dB. When we apply MLG sampling, the model achieves 5.54 and 1.09 dB improvements. When jointly exploiting the two techniques, the model is improved by 13.40 and 3.04 dB on the NVS and CT reconstruction tasks. This evidence clearly exhibits the efficacy of Lineformer and MLG sampling strategy.

Visual Analysis. To intuitively show the effectiveness of the two proposed approaches, we further conduct visual analysis on the scene of bonsai. As depicted in Fig. 7, the baseline model fails to preserve the geometry like the tree branches in the projection and blurs high-frequency details such as the edges of the basin in the CT slice. When successively using LS-MSA and MLG sampling, the model reconstructs more structural contents and fine-grained textures.

Parameter Analysis. We conduct parameter analysis regarding the number of line segments M in Eq. (6) and the patch size S in Eq. (14). Please note that we keep the total number of sampled points and X-rays unchanged for fair comparison. When analyzing one parameter, we fix the other at its optimal value. The results are reported in Tab. 3b and 3c. It is clear that, when using different M and S, our SAX-NeRF stably outperforms the baseline model by over 7.81 and 1.64 dB on NVS and CT reconstruction. This evidence suggests the reliability of our method. The model's performance yields its maximum when $M = 160$ and $S = 4$.

Robustness Analysis. We conduct robustness analysis regarding the number of training projections to compare the performance of different methods when given fewer X-ray projection views. The results are plotted as two line charts in Fig. 8, where the vertical axis is PSNR (in dB performance) and the horizontal axis is the number of training projections. Our SAX-NeRF reliably surpasses SOTA methods by large margins when given different numbers of training projections on both NVS (left) and CT reconstruction (right) tasks. Surprisingly, when using even only 60% of training projections, SAX-NeRF still outperforms other algorithms on the NVS task. These results clearly exhibit the superiority and robustness of our proposed method.

Lineformer *vs.* vanilla Transformer. We replace LS-MSA with G-MSA to conduct comparative experiments. For fair comparison, we keep the model parameters the same by fixing the number of channels and heads. The experimental results show that our Lineformer significantly outperforms vanilla Transformer by 5.30 and 1.28 dB on the NVS and CT reconstruction tasks while only requiring 3.41% of vanilla Transformer's computational complexity. This evidence suggests the efficiency advantage of our Lineformer.

5. Conclusion

In this paper, we focus on studying a core problem in sparse-view X-ray 3D reconstruction, *i.e.*, how to effectively capture the various and complex structures penetrated by X-rays. To this end, we propose a novel framework SAX-NeRF. To model 3D structural dependencies in space, we design a Transformer, Lineformer, as the backbone of SAX-NeRF. Lineformer partitions an X-ray into different line segments and then computes self-attention within each piece of the X-ray. In addition, to extract 2D geometry and contextual representations in projection, we present an MLG ray sampling strategy that contains pixel- and patch-level sampling on the informative foreground regions. Besides, we also collect a larger-scale dataset, X3D, covering wider X-ray application scenarios. Comprehensive experiments on X3D show that SAX-NeRF significantly surpasses SOTA algorithms on the NVS and CT reconstruction tasks.

References

[1] Alaaeldin Ali, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia Neverova, Gabriel Synnaeve, Jakob Verbeek, et al. Xcit: Cross-covariance image transformers. In *NeurIPS*, 2021. 3

[2] Anders H Andersen and Avinash C Kak. Simultaneous algebraic reconstruction technique (sart): a superior implementation of the art algorithm. *Ultrasonic imaging*, 1984. 2, 6, 7

[3] Rushil Anirudh, Hyojin Kim, Jayaraman J Thiagarajan, K Aditya Mohan, Kyle Champley, and Timo Bremer. Lose the views: Limited angle ct reconstruction via implicit sinogram completion. In *CVPR*, 2018. 3

[4] Samuel G Armato III, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, et al. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Medical physics*, 2011. 5

[5] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *ICCV*, 2021. 3

[6] Jonathan T. Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *ICCV*, 2021. 2

[7] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *CVPR*, 2022. 2

[8] Ander Biguri, Manjit Dosanjh, Steven Hancock, and Manuchehr Soleimani. Tigre: a matlab-gpu toolbox for cbct image reconstruction. *Biomedical Physics & Engineering Express*, 2016. 5

[9] JM Boone, N Shah, and TR Nelson. A comprehensive analysis of coefficients for pendant-geometry cone-beam breast computed tomography. *Medical physics*, 2004. 1

[10] John M Boone, Thomas R Nelson, Karen K Lindfors, and J Anthony Seibert. Dedicated breast ct: radiation dose and image quality evaluation. *Radiology*, 2001. 1

[11] Stefan Bruckner and M Eduard Gröller. Instant volume visualization using maximum intensity difference accumulation. In *Computer Graphics Forum*, 2009. 5

[12] Yuanhao Cai, Jing Lin, Xiaowan Hu, Haoqian Wang, Xin Yuan, Yulun Zhang, Radu Timofte, and Luc Van Gool. Mask-guided spectral-wise transformer for efficient hyperspectral image reconstruction. In *CVPR*, 2022. 3

[13] Yuanhao Cai, Jing Lin, Haoqian Wang, Xin Yuan, Henghui Ding, Yulun Zhang, Radu Timofte, and Luc Van Gool. Degradation-aware unfolding half-shuffle transformer for spectral compressive imaging. In *NeurIPS*, 2022.

[14] Yuanhao Cai, Hao Bian, Jing Lin, Haoqian Wang, Radu Timofte, and Yulun Zhang. Retinexformer: One-stage retinex-based transformer for low-light image enhancement. In *ICCV*, 2023. 3

[15] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *ECCV*, 2022. 1, 2, 6, 7

[16] Biao Chen and Ruola Ning. Cone-beam volume ct breast imaging: Feasibility study. *Medical physics*, 2002. 1

[17] Xingyu Chen, Qi Zhang, Xiaoyu Li, Yue Chen, Ying Feng, Xuan Wang, and Jue Wang. Hallucinated neural radiance fields in the wild. In *CVPR*, 2022. 2

[18] Zhang Chen, Zhong Li, Liangchen Song, Lele Chen, Jingyi Yu, Junsong Yuan, and Yi Xu. Neurbf: A neural fields representation with adaptive radial basis functions. In *ICCV*, 2023. 2

[19] Hyungjin Chung, Dohoon Ryu, Michael T McCann, Marc L Klasky, and Jong Chul Ye. Solving 3d inverse problems using pre-trained 2d diffusion models. In *CVPR*, 2023. 3

[20] Allan Macleod Cormack. Representation of a function by its line integrals, with some radiological applications. *Journal of applied physics*, 1963. 1

[21] Allan Macleod Cormack. Representation of a function by its line integrals, with some radiological applications. ii. *Journal of Applied Physics*, 1964. 1

[22] Abril Corona-Figueroa, Jonathan Frawley, Sam Bond-Taylor, Sarath Bethapudi, Hubert PH Shum, and Chris G Willcocks. Mednerf: Medical neural radiance fields for reconstructing 3d-aware ct-projections from a single x-ray. In *International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2022. 2

[23] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 3

[24] Idris A Elbakri and Jeffrey A Fessler. Segmentation-free statistical image reconstruction for polyenergetic x-ray computed tomography with experimental validation. *Physics in Medicine & Biology*, 2003. 1

[25] Lee A Feldkamp, Lloyd C Davis, and James W Kress. Practical cone-beam algorithm. *Josa a*, 1984. 2, 6, 7

[26] Godfrey N Hounsfield. Computerized transverse axial scanning (tomography): Part 1. description of system. *The British journal of radiology*, 1973. 1

[27] Godfrey N Hounsfield. Computed medical imaging. *Science*, 1980. 1

[28] Drew A Hudson and Larry Zitnick. Generative adversarial transformers. In *ICML*, 2021. 3

[29] Yifan Jiang, Shiyu Chang, and Zhangyang Wang. Transgan: Two pure transformers can make one strong gan, and that can scale up. In *NeurIPS*, 2021. 3

[30] Yoni Kasten, Daniel Doktofsky, and Ilya Kovler. End-to-end convolutional neural network for 3d reconstruction of knee bones from bi-planar x-ray images. In *MICCAIW 2020*, 2020. 3

[31] Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[32] Pavol Klacansky. Scientific visualization datasets, 2022. 5

[33] Yiqun Lin, Zhongjin Luo, Wei Zhao, and Xiaomeng Li. Learning deep intensity field for extremely sparse-view cbct reconstruction. In *MICCAI*, 2023. 3

[34] Stephen H Manglos, George M Gagne, Andrzej Krol, F Deaver Thomas, and Rammohan Narayanaswamy. Transmission maximum-likelihood reconstruction with ordered subsets for cone beam ct. *Physics in Medicine & Biology*, 1995. 2

[35] B Mildenhall, PP Srinivasan, M Tancik, JT Barron, R Ramamoorthi, and R Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 1, 2, 6, 7

[36] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM ToG*, 2022. 2, 4

[37] Jinxiao Pan, Tie Zhou, Yan Han, Ming Jiang, et al. Variable weighted ordered subset image reconstruction algorithm. *International Journal of Biomedical Imaging*, 2006. 2

[38] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 5

[39] Philips. Philips research, hamburg, germany. https://teem.sourceforge.net/nrrd/volvis/index.html, 2022. 5

[40] Johann Radon. 1.1 über die bestimmung von funktionen durch ihre integralwerte längs gewisser mannigfaltigkeiten. *Classic papers in modern diagnostic radiology*, 2005. 2

[41] Peter Rautek, Balázs Csébfalvi, Sören Grimm, Stefan Bruckner, and Meister Eduard Gröller. D2vr: High-quality volume rendering of projection-based volumetric data. In *Proceedings of the Eighth Joint Eurographics/IEEE VGTC conference on Visualization*, 2006. 5

[42] Holger R Roth, Le Lu, Amal Farag, Hoo-Chang Shin, Jiamin Liu, Evrim B Turkbey, and Ronald M Summers. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In *MICCAI*, 2015. 5

[43] Darius Rückert, Yuanhao Wang, Rui Li, Ramzi Idoughi, and Wolfgang Heidrich. Neat: Neural adaptive tomography. *TOG*, 2022. 1, 2, 6, 7

[44] Ken Sauer and Charles Bouman. A local update strategy for iterative reconstruction from projections. *TIP*, 1993. 2

[45] William C Scarfe, Allan G Farman, Predag Sukovic, et al. Clinical applications of cone-beam computed tomography in dental practice. *Journal-Canadian Dental Association*, 2006. 1

[46] Emil Y Sidky and Xiaochuan Pan. Image reconstruction in circular cone-beam computed tomography by constrained, total-variation minimization. *Physics in Medicine & Biology*, 2008. 2, 6, 7

[47] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021. 3

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 2, 3

[49] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, and Baining Guo. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In *CVPR*, 2023. 2

[50] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncell. Image quality assessment: from error visibility to structural similarity. *TIP*, 2004. 5

[51] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPS*, 2021. 3

[52] Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. Point-nerf: Point-based neural radiance fields. In *CVPR*, 2022. 2

[53] Xingde Ying, Heng Guo, Kai Ma, Jian Wu, Zhengxin Weng, and Yefeng Zheng. X2ct-gan: reconstructing ct from biplanar x-rays with generative adversarial networks. In *CVPR*, 2019. 3

[54] Lifeng Yu, Yu Zou, Emil Y Sidky, Charles A Pelizzari, Peter Munro, and Xiaochuan Pan. Region of interest reconstruction from truncated data in circular cone-beam ct. *TMI*, 2006. 2

[55] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, 2022. 3

[56] Guangming Zang, Ramzi Idoughi, Rui Li, Peter Wonka, and Wolfgang Heidrich. Intratomo: self-supervised learning-based tomography via sinogram synthesis and prediction. In *CVPR*, 2021. 1, 3, 4, 6, 7

[57] Wojciech Zbijewski, Michel Defrise, Max A Viergever, and Freek J Beekman. Statistical reconstruction for x-ray ct systems with non-continuous detectors. *Physics in Medicine & Biology*, 2006. 2

[58] Nicolas ZCarion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 3

[59] Ruyi Zha, Yanhao Zhang, and Hongdong Li. Naf: neural attenuation fields for sparse-view cbct reconstruction. In *MICCAI*, 2022. 1, 2, 4, 5, 6, 7

[60] Bowen Zhang, Shuyang Gu, Bo Zhang, Jianmin Bao, Dong Chen, Fang Wen, Yong Wang, and Baining Guo. Styleswin: Transformer-based gan for high-resolution image generation. In *CVPR*, 2022. 3

[61] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H.S. Torr, and Li Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *CVPR*, 2021. 3

[62] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H.S. Torr, and Li Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *CVPR*, 2021. 3

[63] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*, 2021. 3