Generalizable Structure-Aware INF: Biplanar-View CT Reconstruction via Disentangled Implicit Neural Field

Bei Huang[0009−0006−3634−4157] and Yuru Pei✉[0000−0001−8520−3509]

School of Intelligence Science and Technology, Key Laboratory of Machine Perception (MOE), State Key Laboratory of General Artificial Intelligence, Peking University, Beijing 100871, China peiyuru@cis.pku.edu.cn

Abstract. Structure-aware CT reconstruction from a single or biplanar X-rays produces patient-specific 3D insights into underlying structures, pushing the radiation hazards during diagnosis and treatment to a minimum. Existing implicit neural fields (INF) methods have shown impressive performance in CT reconstruction, though they rely on multi-view X-rays and conduct subject-specific reconstruction. Additional online adaptation is required to handle novel subjects. In this paper, we present the generalizable structure-aware implicit neural fields (GSA-INF), a unified model that learns a generalizable structure-aware volume prior to INF decoding from sparse-view X-rays. Previous CoderNeRF views the latent code as a holistic shape prior to 3D reconstruction. In contrast, we present a new triplane generative model to learn a generalizable volume prior distribution, where the sampled triplane latent code produces voxel-level representation for INF decoding and CT reconstruction. Moreover, we introduce anatomical structure mask supervision by building a parallel INF-based decoding framework that enhances structure disentanglements when popping up a variety of structures from 2D X-rays. Our approach entails simultaneous INF-based CT reconstruction and volume-prior learning. In the online inference process, we can conditionally reconstruct CT from single- or biplanar-view X-rays and unconditionally generate CTs via sampling in the latent space. GSA-INF demonstrates robust and superior results over the compared methods.

Keywords: Generalizable structure-aware INF · Single-/biplanar-view CT reconstruction · Implicit neural field

1 Introduction

Sparse-view CT reconstruction produces patient-specific 3D morphological information and greatly reduces radiation exposure during diagnosis and treatment. Although numerous methods have emerged to handle single- or biplanar-view CT reconstruction, it remains a major challenge to design a generalizable reconstruction framework that disentangles different types of anatomical structures from input X-rays and accounts for inter-subject shape variations. Compared to

sparse-view-based 3D scene reconstruction in computer vision, this problem is more severely ill-posed due to the need to estimate dense voxels inside anatomical structures, in addition to those on object surfaces.

Implicit neural fields (INF) have shown impressive results in CT reconstruction by solving the inverse volume rendering problem via volumetric deformation and adaptation. The coordinate-based learning methods use the MLP-based mapping function to convert query 3D coordinates to density or attenuation values. However, most existing work conducts subject-specific 3D reconstruction using multi-view X-rays, where additional test-time adaptation is required when confronted with X-rays from novel subjects. To figure out volumes that are specific to an X-ray, image-conditioned CT reconstruction methods use feature projection of the query 3D point and a feed-forward X-ray-to-CT decoder [42, 19]. Considering a query 3D voxel, the image-conditioned models use the same feature embedding for voxels along the sample ray. Thus, the pose embedding and the incorrectly conditioned features for query voxels are insufficient to disambiguate overlapping anatomical structures without 3D prior information.

In order to handle inter-object shape variations, the CodeNeRF [17] integrates the 1D code vectors to learn the NeRF-based shape and texture decoders. The latent coding accounts for a holistic structure prior to 3D reconstruction, but the photometric loss used in code optimization is limited to handling local voxel-level details. The pre-trained generative model [5, 23, 20] has been used to provide volume priors. The diffusion model refined FBP-based CT reconstruction, where the two-stage reconstruction relied on CT initialization and delicately designed regularization in the reverse denoising process. The holistic code guidance and FBP-based initialization for single- or biplanar-view CT reconstruction have limitations in terms of structure disambiguation and cross-dimensional correlation reasoning.

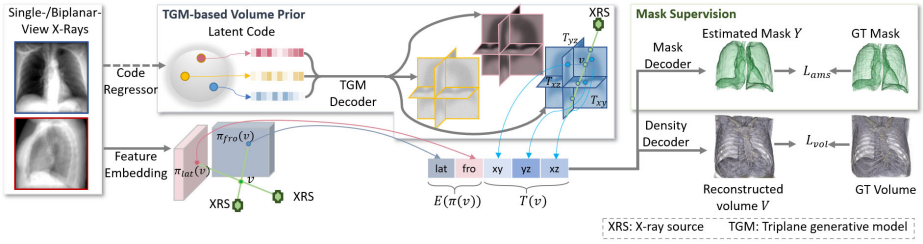

Fig. 1. An overview of our framework. The triplane generative model (TGM) accounts for voxel-level volumetric representation distribution. The TGM decoder receives the latent code as input and produces a triplane-based volumetric representation. This volume prior, along with feature embedding of single- or biplanar-view X-rays, guides CT reconstruction. An optional code regressor is used to initialize the latent code. We build parallel INF-based density and mask decoders to infer voxel density and anatomical structure labels.

In this paper, we propose generalizable structure-aware implicit neural fields (GSA-INF) for single- or biplanar-view CT reconstruction. Concretely, we integrate voxel-level volumetric prior learning and anatomical structure mask supervision into the INF for structural disentanglements and generalized volume reasoning. We use a voxel-level triplane generative model (TGM) instead of the holistic shape prior from latent coding to take into account the distribution of the volumetric representation. TGM entails generalizable CT reconstruction conditioned on input X-rays, as well as unconditioned CT generation by random code sampling in the latent space. We also introduce parallel INF decoders that use structure-aware mask supervision to enhance 3D structure estimation. In single- or biplanar-view CT reconstruction, both the TGM-based volumetric prior and the masked volume supervision fulfill the structure disentanglements and the uncertain voxel inference. We have applied GSA-INF to CT reconstruction in the public LIDC-IDRI datasets. Quantitative and qualitative results have demonstrated robust and superior performances over the compared methods. The main contributions include:

- We present TGM-based voxel-level volumetric prior learning that takes into account the distribution of volumetric representations and allows for both conditional sparse-view CT reconstruction and unconditional CT generation.
- We introduce the mask supervision into the INF-based structure-aware CT reconstruction, which resolves the uncertainty in voxel inference for a variety of 3D anatomical structures from X-rays.
- We have performed extensive experiments on single- or biplanar-view CT reconstruction on publicly available datasets. Comprehensive results have shown the effectiveness and performance gains of the proposed approach over state-of-the-art methods.

2 Related Work

Deep Learning-Based CT Reconstruction. The deep learning-based CT reconstruction framework employs deep neural network-based 2D and 3D image prior distribution for volume generation, greatly relieving the computationally expensive online optimization [26, 39, 35]. The encoder-decoder architecture has been the mainstream cross-modal generator for CT reconstruction [45, 38, 8, 11, 4]. To address the consistent correspondence between CT volumes and projective X-rays, existing work used cross-dimensional feature transformation and duplication to convert 2D image features to volumetric image features [32, 15, 22]. The vector quantification has been used for volumetric coding inference [18]. Manifold approximation-based automated transform [45] and adversarial learning [41] have been used to enhance structural recovery. However, the cascaded CNN-based framework is well known for its memory burden when confronted with high-resolution volumetric images and its struggle with detailed structure recovery.

INF-Based CT Reconstruction. INF has been widely used in scene representation, where a light-weight MLP parameterizes the mapping from the world

coordinates to the voxel values [33, 40, 6]. The INF acts as the continuous representation of 3D space and has been used for CT reconstruction [9, 44, 7]. Neural adaptive tomography (NeAT) introduced the adaptive and hierarchical neural rendering pipeline for CT reconstruction [29]. Vasconcelos et al. [36] studied a Bayesian reformulation of INFs. Reed et al. [27] proposed a limited-view 4D-CT reconstruction method using the INF coupled with a parametric motion field and a differentiable analysis-by-synthesis scheme. In order to refine the INF-based CT reconstruction, the geometrical priors [43], the image prior embedding, and the physics [31] are employed for sparsely sampled measurement embedding. The differentiable rendering formulation [25] and the soft rasterizer [21] were used in self-supervised learning. However, most INF-based CT reconstruction is subject-specific, where the test-time adaptation [34] or retraining is required to adapt the noisy INF-based CT representation to the unseen domain.

Generalizable INF. Image-conditioned INF exploited 2D image feature embedding to generalize the INF and relieve online optimization [42]. The 2D feature embedding combined with low-dimensional volumetric embedding has been used for generalized CT reconstruction [19]. Considering the same feature embedding for voxels along a ray, the above methods lack control over 3D anatomical shapes and structure disentanglements. To overcome this issue, the 3D shape-related code is introduced to disentangle geometry in the INF. NeRF decoder conditioned on the 3D morphable face model (3DMM) has disentangled control of scene appearance and facial actions for face reconstruction [2, 12, 13, 46]. The morphable radiance field (MoRF) extended the NeRF for multi-view facial image synthesis [37]. Unlike scene reconstruction and view synthesis using codes of sparse and deformable point clouds on object surfaces, CT reconstruction needs to fulfill the voxel-wise inference of anatomical structures in 3D space. Existing coded NeRF cannot be directly deployed on voxel-level CT reconstruction.

3 Method

To build the generalizable CT reconstruction model, we propose GSA-INF, a framework that integrates the voxel-level volume prior and masked volume supervision with the INF decoder. Fig. 1 provides an overview of our approach. When given single- or biplanar-view X-rays $I \in \mathcal{I}$, the goal is to estimate the 3D CT image $V \in \mathcal{V}$. We formulate this problem by learning a mapping function f from an X-ray and an associated latent code $z \in \mathcal{Z}$ to the volumetric space $\mathcal{V}$, and each pair (I, z) is associated with a CT volume $V \in \mathcal{V}$ as

$$V = f(I, z), z \in \mathcal{Z}, I \in \mathcal{I}, V \in \mathcal{V}. \tag{1}$$

We use the learned TGM decoder to derive the voxel-level triplane representation T from the sampled latent code z. As the coordinate-based INF, a voxel with 3D coordinates is associated with the combinational feature $F(v) = (E(p), T(v))$, where $p = \pi(v)$ is the associated projective pixel of v on the X-ray feature embedding E when given the projective function π. We formulate the CT images as a set of voxels on a sampled grid with a density value of c and structural labels

of y, which are derived from the parallel INF-based decoders. We optimize the model under masked volume supervision.

Unlike existing image- or code-conditioned INF, our method exploits the voxel-level volume prior and explicitly adapts the 3D representation to the input X-ray in the online inference. Furthermore, our method differs from statistical model-based 3D reconstruction in that the code optimization takes advantage of both 2D X-rays and 3D reconstructed volumes.

3.1 TGM-Based Voxel-Level Volume Prior

We introduce the TGM-based voxel-level volumetric prior, which accounts for volumetric representation distribution and guides CT reconstruction. Unlike the image-conditioned INF, which uses positional embedding to disambiguate the 2D X-ray embedding, the TGM directly derives a voxel-level representation from a latent code. We build the TGM upon the U-Net backbone model [28], which parameterizes the mapping function from the latent space to the voxel-level triplane representation. CodeNeRF[17] also used the code embedding for NeRF learning. However, they did not discriminate the code embedding for local voxels, which impeded the detailed structure reconstruction. In contrast, the proposed TGM decoder embeds the latent code for voxel-level volumetric representation to learn the INF-based mask and density decoders. The TGM not only encodes the volume prior but also addresses voxel representation learning, enhancing the INF-based voxel inference.

Latent Code The low-dimensional latent code accounts for the volumetric prior to facilitate the 3D structure disambiguation and voxel-level attribute inference in terms of a variety of anatomical structures. As to the sparse-view CT reconstruction, the optimal code is required to correspond to the volumes that satisfy the input X-rays in the projective perspective. Note that the code is not unique, even for the conditioned CT reconstruction, considering the ill-posed nature of the inverse problem. As to the unconditional CT generation, the sampled code directly determined volumetric perturbations and deformations regarding the resultant CT volumes. There are several ways to initialize the code values, such as random sampling, statistical average, and code regression.

Random Sampling. Note that each X-ray in the training dataset is associated with a latent code. Under the assumption that the latent codes follow the Gaussian distribution $\mathcal{N}(\mu, \sigma)$, where μ and σ are the mean and variance of the latent codes of the training data, We can randomly sample the latent codes $z = \mu(\mathcal{I}) + \sigma(\mathcal{I}) \odot \epsilon$, and $\epsilon \in \mathcal{N}(0, I)$ for code initialization in conditioned CT reconstruction and unconditional CT generation.

Statistical Average. Previous work of the CodeNeRF [17] relied on the average embedding from the training data as the initial code for 3D reconstruction. In the test-time CT reconstruction, the latent code can be initialized as the statistical average μ of the latent code regarding the training data, where online optimization is used to refine the code specific to the input image.

Code Regressor. As shown in Fig. 1, the ResNet-based code regressor bridges the input X-rays to the latent codes, where the code conditioned on the 2D image promotes triplane decoding specific to the input X-ray and 3D reconstruction. The code regressor initializes the latent code more specific to the input X-ray than the statistical average or random sampling, though it introduces additional model complexities. Generally speaking, the code regressor produces a good starting value, which is feasible to reduce the computational complexity of online code optimization.

TGM-Based Triplane Representation We concatenate the TGM-based triplane representation with the X-ray image embedding as the input for the INF-based density and mask decoders. As to 2D X-ray features, the Resent-based multi-scale k-channel feature embedding has been used in image-conditioned NeRF to enhance the query 3D point via a predefined projective function π. As to voxel v, $E(\pi(v))$ denotes the associated k-dimensional feature vector. 2D feature embedding combined with positional embedding is limited to disambiguating the overlapped tissues, considering all voxels on one sampled ray have the same 2D feature vectors. In this work, the TGM-based triplane encodes the volumetric representation in three orthogonal planes, where arbitrary two planes provides positional information for the 3D query voxel. $(T_{xy}, T_{yz}, T_{xz}) \in \mathbb{R}^{3\times q\times m}$ denotes the bi-linearly interpolated q-channel voxel-level triplane features. The TGM learned from the training data encodes the distribution of triplane representation. The triplane conditioned on the latent code provides volumetric prior to guiding reconstructions of multiple types of anatomical structures.

3.2 Parallel Density and Mask Decoder

In single- or biplanar-view-based CT reconstruction, we need to pop up a variety of anatomical structures, which is a severely ill-posed problem. To distinguish various anatomical structures, we incorporate semantic structure supervision into the voxel inference process. Each voxel in the reconstructed CT is characterized by its density value c and class label $y \in \mathcal{Y}$, where $\mathcal{Y}$ denotes all structural categories. We design parallel INF-based density and structural mask decoders to infer voxel density c and semantic structure labels y from the combinational voxel feature embedding.

Density Decoder. We adopt a density decoder to estimate the voxel density values c from combinational features $F = (E(\pi(v)), T(v))$, which is implemented using the five-layer MLP.

$$c(v) = Relu(MLP_2(Relu(MLP_1(F(v)) + F(v)))). \tag{2}$$

MLP_1 consists of two fully connected layers, and MLP_2 consists of three fully connected layers with a residual connection. $F(v)$ denotes the concatenation of the TGM-based prior volumetric representation and the X-ray embedding. When given voxels sampled on regular grids with a specified resolution, the density decoder enables CT reconstruction.

Mask Decoder. We adopt the INF decoder to infer anatomical structure labels. In this work, we consider the binary mask inference of the left and right lungs. Note that the mask decoder introduces the volumetric mask supervision into learning the TGM-based volumetric representation distribution and the INF-based CT reconstruction. We use the same MLP-based architecture as the density decoder. The additional softmax operator is used to generate the class label y.

$$y(v) = softmax(MLP_2(Relu(MLP_1(F(v)) + F(v))). \tag{3}$$

3.3 Loss

As to the single- or biplanar-view CT reconstruction, we adopt masked volume supervision to optimize the code regressor, the TGM, and the parallel INF-based decodes. The loss function is defined as a linear combination of volume consistency between estimated masked volumes and the ground truth.

$$\mathcal{L} = \mathcal{L}_{vol} + \gamma \mathcal{L}_{ams}. \tag{4}$$

$\mathcal{L}_{vol} = \|V - V_{gt}\|_2^2$, which measures the distance from the estimated volume V to the ground truth CT volume V_{gt}. We use the Dice similarity coefficient to evaluate the distance between estimated anatomical structure mask Y and the ground truth Y_{gt}, and $\mathcal{L}_{ams} = 1 - 2\frac{|Y \cap Y_{gt}|}{|Y + Y_{gt}|}$. Unlike the 2D supervision used in NeRF-based sparse-view 3D reconstruction, we step back to 3D supervision in that the volume rendering-based consistency on the 2D plane is limited to reasoning dense voxels of a variety of anatomical structures. The hyperparameter γ is used to balance the volume density supervision and mask supervision.

3.4 Online Inference

When it comes to single- or biplanar-view CT reconstruction, the proposed model initializes the latent code z using the code regressor, which is used to generate an X-ray conditioned triplane. The combinational feature embedding F is fed to the density decoder to reconstruct CTs. Random sampling and statistical averages cannot guarantee consistency between the latent code and the target volume. Even the code regressor conditioned on the input X-ray cannot ensure the optimal latent code. We employ test-time code optimization to refine the latent code and enhance the volumetric reconstruction (Alg. 1).

Code Optimization When given 2D X-ray images I, we first initialize the latent code z. The proposed GSA-INF produces the CT volume V via the learned density decoder. In the online testing process, the projection of the reconstructed volume is required to be consistent with the input X-rays. As shown in Fig. 2, we optimize z by minimizing l_2 norm-based image distance between the rendered X-ray and the input. The latent code is iteratively updated as

$$z = z + \kappa \nabla \|I - I_r\|_2^2, \tag{5}$$

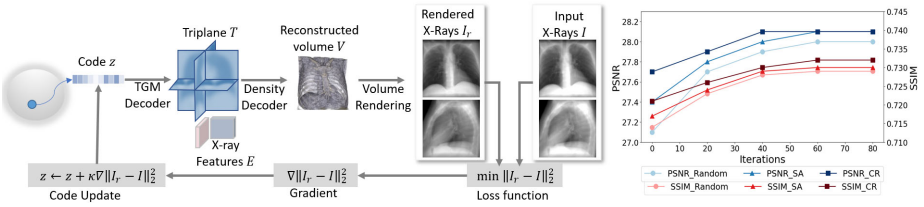

Fig. 2. Left: online code optimization process. Right: CT reconstruction accuracy during online code optimization. The latent codes are initialized by random sampling, statistical average (SA), and the code regressor (CR).

where I_r denotes the rendered X-ray from the reconstructed CT V. The predefined device-specific parameters, such as the object-image distance and the source-object distance, determine the volume renderer. To render X-rays, we use the drrsuite [10], where slices perpendicular to the projection direction are scaled and accumulated together for the synthetic X-rays. We normalize the rendered X-ray image to $[0, 1]$.

Algorithm 1 Test-time code optimization.

Input: Single-/biplanar-view X-rays I, learned GSA-INF, maximum iteration number M, X-ray consistency threshold η;
Output: Latent code z;

$Iter = 0$;
while $iter < M$ and $r > \eta$ **do**
  Initialize the latent code z from the input X-ray I via the code regressor;
  Generate the triplane representation from z using the TGM decoder;
  for Each sampled voxel of volume V **do**
    Compute density via the INF-based density decoder;
  end for
  Generate the rendered X-rays I_r via the volume renderer;
  Compute the loss $r = \|I - I_r\|_2^2$;
  Update the latent code z via (5);
  $Iter$++;
end while

Unconditional CT Generation Random sampling in the latent code space entails unconditional CT generation. Note that the density decoders take combinational features, i.e., the latent code-conditioned triplane and the X-ray-conditioned feature embedding, as input. Aside from randomly sampled latent code, we set the X-ray embedding as the average feature embedding of the training data or randomly selected reference X-rays, where volumetric CTs are generated using the combinational features.

4 Experiments

4.1 Dataset

We conduct experiments on the public chest CT dataset of the LIDC-IDRI [1] for benchmarking with previous work. The LIDC-IDRI dataset contains 1,018 chest CT scans and synthetic X-rays, with a train/test split of 912/102. The chest CTs have been cropped and re-sampled to a resolution of $128 \times 128 \times 128$ with an isotropic voxel size of $2.5mm \times 2.5mm \times 2.5mm$. Each CT scan has associated orthogonal view X-rays with a resolution of 128×128, which are generated by the drrsuite toolkit [10]. We use the left and right lung annotations provided by the LUNA16 dataset [30], which contains 888 annotated CTs with training/testing split of 800/88. The training dataset contains 912 scans in LIDC-IDRI, of which 800 have lung annotations from the LUNA16 dataset.

4.2 Implemental Details

We utilize ResNet-34 [14] for feature embedding of 2D lateral and frontal X-rays. The X-ray feature embedding channel number k is set to 1024. We use five-scale feature maps with a resolution ranging from 4×4 to 64×64, with channel numbers of 512, 256, 128, 64, and 64, respectively. The TGM decoder is built upon U-Net [28]. We use ResNet-34, followed by a two-layer MLP, as the optional code regressor for a 512-dimensional latent code. The triplane features have a resolution of $m = 64 \times 64$ and the channel number $q = 354$. Both the density and mask decoder are built upon a five-layer MLP. We conduct the spatial sampling with approx. $250k$ voxels from each chest CT. The hyper-parameter γ in the loss function (4) is set to 1e-2. κ used in the test-time code optimization (5) is set to 1e-2. We implement the proposed GSA-INF using PyTorch on a machine with a GTX 3090 GPU. We use the Adam algorithm with a learning rate of 1e-4. The learning rates decay by 0.5 every 25 epochs after 100 epochs. Each mini-batch contains $250k$ sampled voxels from a CT volume in the X-ray-based volume recovery. The training takes 36 hours, with 150 epochs. The online inference by the single- and biplanar X-rays takes an average of 5.3 seconds and 7.2 seconds in the online testing process for CT reconstruction. Online optimization takes an average of 50 seconds.

4.3 Single-/Biplanar-View CT Reconstruction

In this section, we conduct evaluations for single- and biplanar-view CT reconstruction on the LIDC-IDRI dataset. Large-scale lung shape deformations are involved in CT scanning with different respiratory phases. Chest CT reconstruction from sparse-view X-rays is a challenging issue, where a variety of structures, such as the rib cage and lungs, need to be popped up from 2D X-rays.

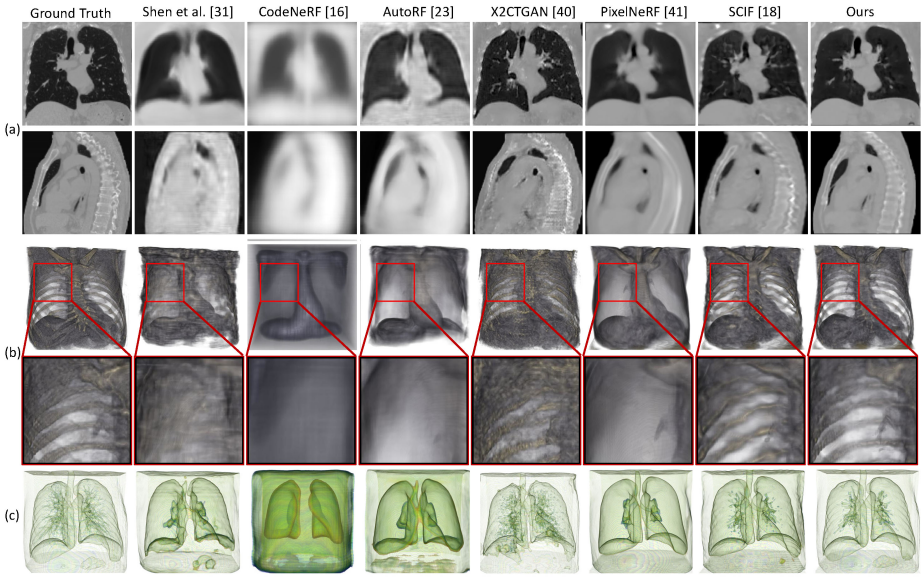

Fig. 3. CT reconstruction from biplanar-view X-rays by compared methods. (a) Sampled sagittal and coronal slices. (b) Volume rendering with visible rib cages. (c) Volume rendering with visible lungs and tracheas.

Evaluation Protocol and Metrics Given input X-rays from the testing scans, we conduct the triplane decoding from the latent codes initialized via the code regressor, random sampling, or the statistical average. We combine the triplane decoding with the X-ray embedding to guide the CT reconstruction. Test-time code fine-tuning with respect to the input X-rays yields the final CTs. We use three kinds of image quality metrics, including the NCC, the peak signal-to-noise ratio (PSNR), and structural similarity (SSIM), to evaluate reconstructed CTs.

Table 1. CT reconstruction accuracies of compared methods.

	Single			Biplanar		
	PSNR	SSIM	NCC	PSNR	SSIM	NCC
Shen et al. [32]	22.6±1.24	0.457±0.061	0.623±0.052	25.0±0.99	0.538±0.058	0.801±0.049
Henzler et al. [15]	22.4±1.43	0.438±0.076	0.598±0.032	24.0±1.45	0.534±0.090	0.723±0.061
X2CTGAN [41]	23.1±0.02	0.525±0.004	0.692±0.033	26.2±0.13	0.656±0.008	0.912±0.019
CodeNeRF [17]	21.8±0.98	0.428±0.090	0.588±0.012	24.1±1.12	0.525±0.086	0.715±0.029
AutoRF [24]	22.7±0.37	0.471±0.032	0.613±0.045	25.2±0.45	0.542±0.012	0.799±0.021
PixelNeRF [42]	23.8±0.76	0.555±0.006	0.674±0.027	27.2±0.63	0.712±0.003	0.921±0.011
SCIF [19]	24.2±1.22	0.599±0.004	0.691±0.031	27.4±1.04	0.714±0.004	0.931±0.013
Ours	**24.4±0.57**	**0.604±0.031**	**0.717±0.009**	**28.2±0.32**	**0.735±0.012**	**0.952±0.010**

Comparison to the state-of-the-art Table 1 and Fig. 3 report the comparison of the proposed GSA-INF with state-of-the-art methods. Our approach achieves the best performances in both tasks, indicating the plausible image quality and reconstruction accuracy. Previous work [32, 15, 41] used 2D/3D cascaded CNN, where features had to be duplicated or reshaped to convert 2D feature embedding to 3D volumetric feature embedding. The volumetric convolutions are known to be memory-intensive, which constrains the resolution of CT volumes. Moreover, the reconstructed volumes tend to be blurry, considering the relatively large perceptual fields. X2CTGAN [41] used the adversarial learning to add structural details, though there is no guarantee that the recovered details consistent with the ground truth. For a fair comparison, we replace the 2D supervision with volume supervision in the compared INF-based methods [17, 24, 42, 19]. Generalized CT reconstruction can use the image-conditioned INF, but they have limitations in extracting a variety of structures from the input X-ray embedding without volumetric prior guidance. CodeNeRF and AutoRF [17, 24] introduce latent code embedding in 3D reconstruction. The AutoRF [24] introduces the code regressor to infer latent code, which provides the latent codes conditioned on the input images. The code-based methods [17, 24] employ code embedding as a holistic representation without addressing voxel-level feature embedding, which limits its capacity to discover local structural details of underlying anatomical structures. The resultant CTs via [17, 24] are extremely blurry. The 2D object masking [24] is feasible to constrain object shapes, though it is limited to disentangle underlying anatomical structures. The proposed GSA-INF benefits from the TGM-based voxel-level volume prior and the mask supervision, showing performance gains over compared methods. For instance, our method achieves PSNR gains of 0.2, 0.6, 1.7, 2.6, and 1.3 over SCIF, PixelNeRF, AutoRF, CodeNeRF, and X2CTGAN, respectively.

	CR	CO	MS	PSNR	SSIM	NCC
M_1	✓			27.5±0.51	0.716±0.028	0.932±0.005
M_2		✓		28.0±0.67	0.729±0.029	0.949±0.003
M_3	✓	✓		28.1±0.29	0.731±0.022	0.951±0.002
M_4	✓		✓	27.7±0.44	0.725±0.035	0.936±0.010
M_5		✓	✓	28.1±0.27	0.731±0.014	0.950±0.008
M_6	✓	✓	✓	**28.2±0.32**	**0.735±0.012**	**0.952±0.010**

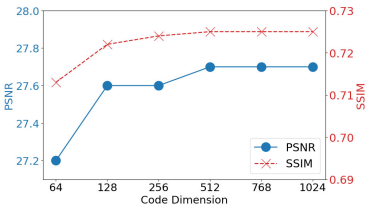

Fig. 4. Left: Ablation studies on the code regressor (CR), online code optimization (CO), and the mask supervision (MS). Right: CT reconstruction accuracy with increasing dimensionalities of the latent codes.

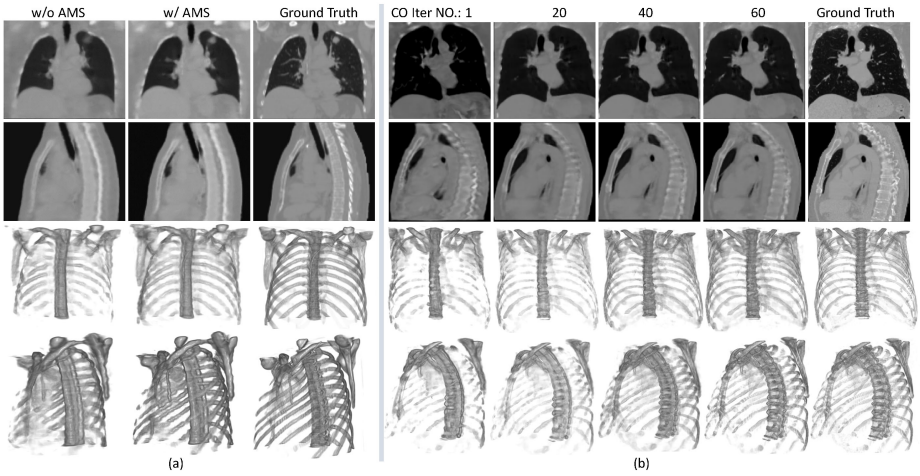

Fig. 5. (a) Effectiveness of anatomical mask supervision (AMS). (b) Reconstructed CTs with increasing iteration numbers in test-time code optimization. From top to bottom: Sampled coronal and sagittal slices, volume rendering with visible rib cages.

4.4 Ablation Study

As shown in Fig. 4-Left, we perform the ablation study and evaluate the effectiveness of the mask supervision, the test-time code optimization, the TGM-based volume prior, and the code regressor on CT reconstruction.

Mask Supervision We build parallel INF-based decoders to introduce structural mask supervision into the model optimization. Fig. 5 (a) shows the effectiveness of the mask supervision in the biplanar-view CT reconstruction. We note that the mask supervision has positive effects on the reconstruction of a variety of anatomical structures, even though we only used the lung masks for supervision. For instance, in the reconstructed CTs, we observe rib cages with clear rib arrangements that are consistent with the ground truth.

Test-Time Code Optimization We conduct the code optimization to refine the latent TGM code to adapt to the input X-ray, which is further used to generate triplane volume prior to CT reconstruction. Fig. 5 (b) and Fig. 6 show refined CT volumes and quantitative CT reconstruction accuracy with increasing numbers of iterations. Note that even when given the code regressor to predict the TGM code from input X-rays, the online code optimization is feasible to improve the CT reconstruction performance. For instance, the PSNR and the NCC increases by 0.5-0.6 and 0.019-0.026 when given the online optimization. We think the reason is that X-ray embedding alone is limited to recovering 3D structures with similar accumulative projections. On the other hand, the

iteratively updated volumes contribute to the code optimization that is specific to a variety of anatomical structures.

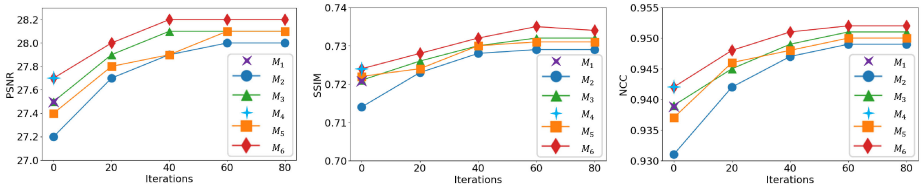

Fig. 6. CT reconstruction accuracy via variant models in the ablation study with increasing numbers of iterations in the test-time code optimization.

TGM-Based Prior We introduce TGM to generate voxel-level volume prior to CT reconstruction. Instead of the global code embedding used for NeRF-based 3D reconstruction [17, 24], the TGM maps the latent code for voxel-level representation and attribute inference. Table 1 shows the performance gains of TGM-based prior over existing image- and code-conditioned methods. For instance, in single- and biplanar-view CT reconstructions, the PSNR improves by 0.6-2.6 and 1.0-4.1 over the PixelNeRF, AutoRF [24], and CodeNeRF [17].

Code Regressor We provide an optional code regressor to infer the latent code from input X-rays. Note that the code regressor is optional because the online code optimization can fine-tune the code conditioned on the input X-rays. In experiments, we observe the positive effects of the code regressor in CT reconstruction combined with online code optimization as shown in Fig. 4-Left. From a numerical optimization perspective, good starting values are beneficial for fast convergence as shown in Fig. 2.

4.5 Parameter Analysis

Code Dimensionality We have introduced a latent code-based TGM decoder to generate volumetric prior and guide CT reconstruction. We can use low-dimensional code to model the distribution of volumetric representations. Fig. 4-Right reports CT reconstruction performances with increasing code dimensionalities. We note that the performance reaches a plateau at 512.

Code Initialization Fig. 2 shows the iterative optimization of the latent codes initialized by random sampling, statistical average, and the ResNet-based code regressor. It is interesting to note that the test-time code optimization converges with all three types of code initialization. We observe that the codes initialized via the code regressor converge faster than those initialized by random sampling

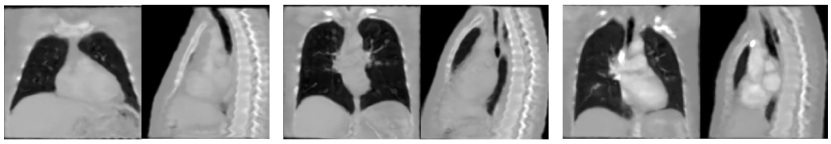

Fig. 7. Sampled coronal and sagittal slices of unconditionally generated CTs.

or the statistical average, which can be ascribed to a good starting values being preferable in the nonlinear optimization.

4.6 Unconditioned CT Generation

Fig. 7 shows sampled sagittal and coronal slices of generated CTs. The proposed GSA-INF is feasible to generate a large variety of morphologies by random code sampling in the latent space. We use the generation metrics of Fréchet Inception Distance (FID) [16] and Kernel Inception Distance (KID) [3]. The testing dataset of LIDC-IDRI serves as the reference image set for the metrics. The proposed method achieves a KID of 0.05 and an FID of 43.2, respectively.

5 Conclusion

In this paper, we propose GSA-INF, which integrates the TGM-based volumetric prior and INF-based representation through a single stage learning paradigm under masked volume supervision. It addresses generalizable single-/bipalnar-view CT reconstruction, overcoming the limitations in existing work where the subject-specific INF relies on dense observations and the prior code embedding ignores voxel-level local detail reconstruction. The GSA-INF shows improvement in performance for robust volumetric image generation. It works well at structural disentanglements when popping up anatomical structures from 2D X-rays.

Limitations and Future Work Our current work relies on ground truth masked volumes during training under the observation that 2D supervision alone is insufficient to adapt the INF to extremely sparse 2D X-rays. Future work may explore few-shot 3D supervision or pre-trained model-generated labels to relieve 3D data collection and annotation. Additionally, the latent codes are sampled under a Gaussian distribution determined by the training data, which affects the fidelity of the generated CTs. A better latent space distribution or sampling scheme is required to address this problem.

Acknowledgments. This work was supported in part by National Natural Science Foundation of China under Grant 62272011 and 61876008, Beijing Natural Science Foundation 7232337.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Armato, S.G., Mclennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., Zhao, B., Aberle, D., Henschke, C.I., Hoffman, E.A., Kazerooni, E.A., MacMahon, H., van Beek, E.J.R., Yankelevitz, D., Biancardi, A.M., Bland, P.H., Brown, M.S., Engelmann, R.M., Laderach, G.E., Max, D., Pais, R.C., Qing, D.P.Y., Roberts, R.Y., Smith, A.R., Starkey, A., Batra, P., Caligiuri, P., Farooqi, A.O., Gladish, G.W., Jude, C.M., Munden, R., Petkovska, I., Quint, L.E., Schwartz, L.H., Sundaram, B., Dodd, L.E., Fenimore, C., Gur, D., Petrick, N., Freymann, J.B., Kirby, J., Hughes, B., Casteele, A.V., Gupte, S., Sallam, M., Heath, M.D., Kuhn, M.H., Dharaiya, E., Burns, R., Fryd, D., Salganicoff, M., Anand, V., Shreter, U., Vastagh, S., Croft, B.Y., Clarke, L.P.: The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. Medical physics **38 2**, 915–31 (2011)
2. Athar, S., Shu, Z., Samaras, D.: Flame-in-nerf : Neural control of radiance fields for free view face animation. ArXiv **abs/2108.04913** (2021)
3. Binkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying mmd gans. ArXiv **abs/1801.01401** (2018)
4. Chen, H., Zhang, Y., Chen, Y., Zhang, J., Zhang, W., Sun, H., Lv, Y., Liao, P., Zhou, J., Wang, G.: Learn: Learned experts assessment-based reconstruction network for sparse-data ct. IEEE Trans. Med. Imag. **37**, 1333–1347 (2018)
5. Chen, H., Gu, J., Chen, A., Tian, W., Tu, Z., Liu, L., Su, H.: Single-stage diffusion nerf: A unified approach to 3d generation and reconstruction. IEEE/CVF International Conference on Computer Vision (ICCV) pp. 2416–2425 (2023)
6. Chen, Z., Zhang, H.: Learning implicit fields for generative shape modeling. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 5932–5941 (2019)
7. Corona-Figueroa, A., Frawley, J., Bond-Taylor, S., Bethapudi, S., Shum, H.P.H., Willcocks, C.G.: Mednerf: Medical neural radiance fields for reconstructing 3d-aware ct-projections from a single x-ray. 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) pp. 3843–3848 (2022)
8. Eyndhoven, G.V., Batenburg, K.J., Kazantsev, D., Nieuwenhove, V.V., Lee, P., Dobson, K., Sijbers, J.: An iterative ct reconstruction algorithm for fast fluid flow imaging. IEEE Trans. Image Process **24**, 4446–4458 (2015)
9. Fang, Y., Mei, L., Li, C., Liu, Y., Wang, W., Cui, Z., Shen, D.: Snaf: Sparse-view cbct reconstruction with neural attenuation fields. ArXiv **abs/2211.17048** (2022)
10. Folkerts, M.M.: drrsuite. In: https://code.google.com/archive/p/drrsuite/
11. Fu, J., Dong, J., Zhao, F.: A deep learning reconstruction framework for differential phase-contrast computed tomography with incomplete data. IEEE Trans. Image Process **29**, 2190–2202 (2020)
12. Gafni, G., Thies, J., Zollhofer, M., Nießner, M.: Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 8645–8654 (2021)
13. Gao, C., Shih, Y., Lai, W.S., Liang, C.K., Huang, J.B.: Portrait neural radiance fields from a single image. ArXiv **abs/2012.05903** (2020)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)

15. Henzler, P., Rasche, V., Ropinski, T., Ritschel, T.: Single-image tomography: 3d volumes from 2d cranial x-rays. Computer Graphics Forum **37** (2018)
16. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Neural Information Processing Systems (2017)
17. Jang, W.J., de Agapito, L.: Codenerf: Disentangled neural radiance fields for object categories. IEEE/CVF International Conference on Computer Vision (ICCV) pp. 12929–12938 (2021)
18. Jiang, Y., Pei, Y., Li, P., Zhang, Y., Guo, Y., Fan, Y., Chen, G., Dai, F., Xu, T., Yuan, X., yan Zha, H.: Craniofacial volumetric image estimation from a lateral cephalogram using cross-dimensional discrete embedding mapping. IEEE Transactions on Computational Imaging **8**, 972–985 (2022)
19. Jiang, Y., Yuan, X., Pei, Y.: Spatially-consistent implicit volumetric function for uni- and bi-planar x-ray-based computed tomography reconstruction. IEEE 20th International Symposium on Biomedical Imaging (ISBI) pp. 1–5 (2023)
20. Liu, J., Anirudh, R., Thiagarajan, J.J., He, S., Mohan, K.A., Kamilov, U.S., Kim, H.: Dolce: A model-based probabilistic diffusion framework for limited-angle ct reconstruction. IEEE/CVF International Conference on Computer Vision (ICCV) pp. 10464–10474 (2022)
21. Liu, S., Li, T., Chen, W., Li, H.: Soft rasterizer: A differentiable renderer for image-based 3d reasoning. 2019 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 7707–7716 (2019)
22. Montoya, J., Zhang, C., Li, K., Chen, G.H.: Volumetric scout ct images reconstructed from conventional two-view radiograph localizers using deep learning. In: Physics of Medical Imaging. vol. 10948 (2019)
23. Muller, N., Siddiqui, Y., Porzi, L., Bulò, S.R., Kontschieder, P., Nießner, M.: Diffrf: Rendering-guided 3d radiance field diffusion. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 4328–4338 (2022)
24. Müller, N., Simonelli, A., Porzi, L., Bulò, S.R., Nießner, M., Kontschieder, P.: Autorf: Learning 3d object radiance fields from single view observations. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 3961–3970 (2022)
25. Niemeyer, M., Mescheder, L.M., Oechsle, M., Geiger, A.: Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 3501–3512 (2020)
26. Preiswerk, F., Toews, M., Cheng, C.C., yuan George Chiou, J., Mei, C.S., Schaefer, L., Hoge, W.S., Schwartz, B.M., Panych, L., Madore, B.: Hybrid mri ultrasound acquisitions, and scannerless realtime imaging. Magnetic Resonance in Medicine **78**, 897¨C908 (2017)
27. Reed, A.W., Kim, H., Anirudh, R., Mohan, K.A., Champley, K.M., Kang, J., Jayasuriya, S.: Dynamic ct reconstruction from limited views with implicit neural representations and parametric motion fields. IEEE/CVF International Conference on Computer Vision (ICCV) pp. 2238–2248 (2021)
28. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. ArXiv **abs/1505.04597** (2015)
29. Rückert, D., Wang, Y., Li, R., Idoughi, R., Heidrich, W.: Neat: Neural adaptive tomography. ACM Trans. Graph. **41**, 55:1–55:13 (2022)
30. Setio, A.A.A., et al.: Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The luna16 challenge. Medical Image Analysis **42**, 1–13 (2016)

31. Shen, L., Pauly, J.M., Xing, L.: Nerp: Implicit neural representation learning with prior embedding for sparsely sampled image reconstruction. IEEE transactions on neural networks and learning systems (2021)
32. Shen, L., Zhao, W., Xing, L.: Patient-specific reconstruction of volumetric computed tomography images from a single projection view via deep learning. Nat. Biomed.Eng. **3**, 880 – 888 (2019)
33. Sitzmann, V., Zollhoefer, M., Wetzstein, G.: Scene representation networks: Continuous 3d-structure-aware neural scene representations. ArXiv **abs/1906.01618** (2019)
34. Song, B., Shen, L., Xing, L.: Piner: Prior-informed implicit neural representation learning for test-time adaptation in sparse-view ct reconstruction. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) pp. 1928–1937 (2023)
35. Syben, C., Michen, M., Stimpel, B., Seitz, S., Ploner, S.B., Maier, A.K.: Pyro-nn: Python reconstruction operators in neural networks. Medical Physics **46**, 5110 – 5115 (2019)
36. Vasconcelos, F., He, B., Singh, N., Teh, Y.W.: Uncertainr: Uncertainty quantification of end-to-end implicit neural representations for computed tomography. ArXiv **abs/2202.10847** (2022)
37. Wang, D., Chandran, P., Zoss, G., Bradley, D., Gotardo, P.F.U.: Morf: Morphable radiance fields for multiview neural head modeling. ACM SIGGRAPH 2022 Conference Proceedings (2022)
38. Wu, Y., Ma, Y., Capaldi, D.P.I., Liu, J., Zhao, W., Du, J., Xing, L.: Incorporating prior knowledge via volumetric deep residual network to optimize the reconstruction of sparsely sampled mri. Magnetic resonance imaging (2019)
39. Würfl, T., Hoffmann, M., Christlein, V., Breininger, K., Huang, Y., Unberath, M., Maier, A.K.: Deep learning computed tomography: Learning projection-domain weights from image domain in limited angle problems. IEEE Trans. Med. Imag. **37**, 1454–1463 (2018)
40. Xu, Q., Wang, W., Ceylan, D., Mech, R., Neumann, U.: Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. In: NeurIPS (2019)
41. Ying, X., Guo, H., Ma, K., Wu, J.Y., Weng, Z., Zheng, Y.: X2ct-gan: Reconstructing ct from biplanar x-rays with generative adversarial networks. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 10611–10620 (2019)
42. Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelnerf: Neural radiance fields from one or few images. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
43. Zang, G., Idoughi, R., Li, R., Wonka, P., Heidrich, W.: Intratomo: Self-supervised learning-based tomography via sinogram synthesis and prediction. IEEE/CVF International Conference on Computer Vision (ICCV) pp. 1940–1950 (2021)
44. Zha, R., Zhang, Y., Li, H.: Naf: Neural attenuation fields for sparse-view cbct reconstruction. ArXiv **abs/2209.14540** (2022)
45. Zhu, B., Liu, J.Z., Rosen, B.R., Rosen, M.S.: Image reconstruction by domain-transform manifold learning. Nature **555**, 487–492 (2018)
46. Zhuang, Y., Zhu, H., Sun, X., Cao, X.: Mofanerf: Morphable facial neural radiance field. In: European Conference on Computer Vision (2021)