Research Article | Open Access

Intelligence & Robotics

Liu *et al. Intell. Robot.* **2026**, *6*(4), 891-918 **DOI:**10.20517/ir.2026.39

AME-Net++: adaptive mixture-of-experts for intelligent intraoperative 3D reconstruction

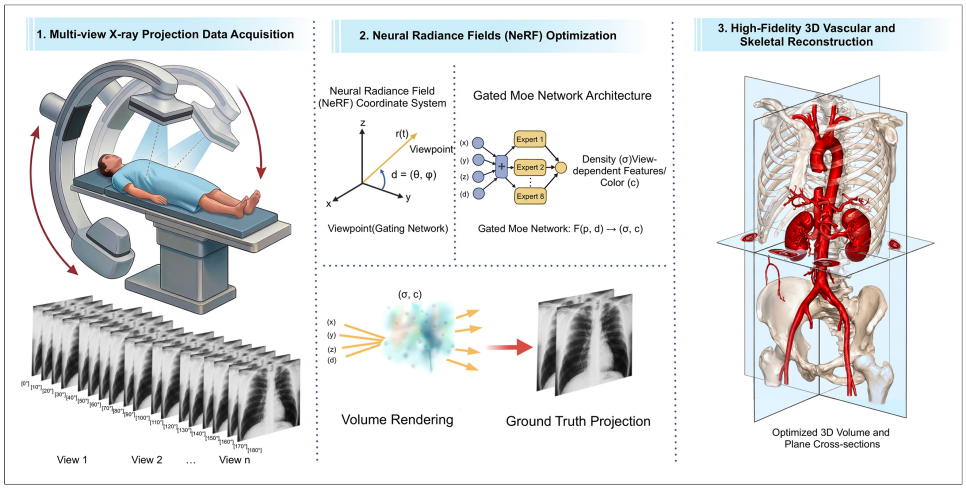
Shi-Qi Liu[1,2], You-Jun Zhang[1], Xiao-Liang Xie[2], Lin-Sen Zhang[2], Xiao-Hu Zhou[2], Xi-Yao Ma[2], Zeng-Guang Hou[2], Xin-Kai Qu[1]

Keywords:
Intraoperative 3D reconstruction, X-ray imaging, adaptive mixture-of-experts learning, multi-scale loss, sparse-view reconstruction

Citation: Liu, S. Q.; Zhang, Y. J.; Xie, X. L.; Zhang, L. S.; Zhou, X. H.; Ma, X. Y.; Hou, Z. G.; Qu, X. K. AME-Net++: adaptive mixture-of-experts for intelligent intraoperative 3D reconstruction. *Intell. Robot.* **2026**, *6*(4), 891-918. https://dx.doi.org/10.20517/ir.2026.39

Received: 10 Jul 2026 **First Decision:** 11 Aug 2026 **Revised:** 24 Sep 2026 **Accepted:** 28 Oct 2026 **Published:** 5 Nov 2026

Academic Editor: Simon X. Yang **Copy Editor:** Pei-Yun Wang **Production Editor:** Pei-Yun Wang



Abstract

In minimally invasive vascular interventions, intraoperative 3D vascular reconstruction and accurate spatial characterization are crucial for safe and precise robotic surgical navigation. However, unlike conventional supervised learning tasks that rely on a fixed training/testing split, projection-based neural reconstruction aims to establish a patient-specific mapping between sparse X-ray observations and the corresponding 3D spatial representation. In this paper, we propose AME-Net++, an Adaptive Mixture-of-Experts Network that integrates feature perception, adaptive expert collaboration, and multi-scale optimization for 3D structural reconstruction from intraoperative multi-view X-ray projections. Specifically, AME-Net++ employs an adaptive feature-preservation strategy to suppress projection noise while preserving the geometric continuity of vascular structures. To address heterogeneous tissue attenuation and region-dependent structural variations, we design a parallel mixture-of-experts architecture with a learnable routing mechanism that dynamically assigns local features to appropriate expert subnetworks for region-adaptive, fine-grained representation learning. In addition, we propose a multi-scale information-aware composite loss function to jointly optimize global structural fidelity, local density consistency, and edge-detail preservation. Experimental results on both public datasets

[1]Department of Cardiology, Huadong Hospital Affiliated to Fudan University, Shanghai 200040, China.
[2]State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China.

Correspondence to: Prof. Xin-Kai Qu, Department of Cardiology, Huadong Hospital Affiliated to Fudan University, Shanghai 200040, China. E-mail: quxinkai@fudan.edu.cn

and our self-built animal interventional vascular dataset demonstrate that AME-Net++ achieves strong performance in both 2D projection synthesis and 3D voxel reconstruction tasks. Additional validation on this dataset shows that the reconstructed vascular centerline has an average error of 1.5 ± 0.67 mm, indicating that the proposed method can preserve the major vascular trajectory with clinically meaningful spatial accuracy. These results validate the effectiveness of the proposed reconstruction method and indicate its potential for 3D navigation in vascular interventional procedures.

1. INTRODUCTION

Benefiting from high operational precision and stability, robotic endovascular interventional systems have demonstrated growing clinical value in modern medical practice. However, autonomous navigation in these systems still relies heavily on single-view X-ray projections for anatomical perception and device path planning. In complex vascular environments, the limited spatial information provided by 2D projections hinders the accurate perception of vascular topology and spatial relationships, thereby increasing procedural complexity and potentially compromising patient safety. Therefore, achieving reliable intraoperative 3D perception remains a critical challenge in developing intelligent robotic endovascular interventional systems.

The mainstream approach provides 3D navigation guidance by registering preoperative computed tomography angiography (CTA) with intraoperative digital subtraction angiography (DSA) images[1]. Although this strategy offers valuable 3D anatomical information, it still has several inherent limitations, such as a strong dependence on accurate cross-modal anatomical correspondence, high computational costs, and performance degradation caused by respiratory motion, cardiac pulsation, or soft-tissue deformation. These issues substantially limit the practicality of registration-based 3D guidance methods for autonomous robotic navigation.

To avoid the limitations of cross-modal registration, direct 3D reconstruction from projection data has emerged as a promising alternative. Existing methods mainly include analytical, iterative, and learning-based approaches. Classical analytical algorithms such as Feldkamp–Davis–Kress (FDK) algorithm are highly efficient under ideal acquisition conditions[2], but their performance degrades significantly in ill-posed settings such as sparse-view reconstruction. Iterative methods improve robustness by incorporating prior regularization[3], but often incur considerable computational and memory overhead. More recently, neural field methods have shown impressive reconstruction capability in tomographic imaging[4,5]. However, applying standard neural radiance fields (NeRF) frameworks directly to computed tomography (CT) reconstruction remains challenging due to their limited ability to model heterogeneous anatomical structures and their underutilization of informative cues embedded in projection images.

NeRF-style reconstruction methods differ from conventional supervised learning methods. The objective is not to train a universal model on a predefined training set and then evaluate its cross-sample prediction performance on a separate testing set. Instead, the neural representation establishes a mapping between projection observations and the corresponding 3D spatial field for each scene or patient. The reconstructed 3D result reflects the representation and reconstruction capability of the method under sparse-view observations.

Although recent studies in general 3D scene reconstruction, such as indoor-scene or furniture reconstruction, have shown that high-quality reconstruction can sometimes be achieved from very few views, it remains unclear whether this conclusion can be directly transferred to medical organ and vascular reconstruction. Compared with natural scenes, medical reconstruction involves more complex attenuation mechanisms, weaker tissue contrast, overlapping anatomical structures, and stricter requirements for both

geometric and density accuracy. Therefore, reliable 3D reconstruction of organs or complex vascular structures from only a few X-ray projections remains a highly ill-posed problem and requires further systematic investigation.

To address the limited applicability of cross-modal registration methods in dynamic intraoperative environments, as well as the limited ability of existing reconstruction methods to model complex vascular structures, we propose an intraoperative 3D reconstruction framework for robotic endovascular intervention, exploring a direct 3D perception paradigm independent of preoperative prior registration. The proposed framework directly reconstructs 3D vascular structures from sequential multi-view intraoperative X-ray projections, aiming to build a high-precision 3D visual representation for autonomous robotic navigation and provide a reliable visual foundation for anatomical perception, path planning, and safe operation in complex vascular environments.

In this work, AME-Net++ is positioned as a reconstruction-quality-oriented neural representation framework. For complex full-volume reconstruction involving multiple anatomical structures, the current optimization process may require substantial computational time and is more suitable for offline or retrospective analysis. For vascular-only reconstruction, the target structure is sparse and concentrated, and the reconstruction process can be completed within a shorter time scale, suggesting potential utility for intraoperative vascular analysis and navigation support.

The main contributions of this work are summarized as follows:
• An adaptive mixture-of-experts reconstruction framework, termed AME-Net++, is proposed for intraoperative vascular structure reconstruction. Through a parallel expert pool and a learnable routing mechanism, it enables adaptive modeling of different tissue types and geometric regions, improving the representation capability of complex heterogeneous vascular structures.
• A multi-scale information-aware weighted composite loss is designed, in which projection-level entropy weighting, local information density, and multi-resolution supervision are jointly incorporated to improve the reconstruction quality of both structural regions and fine details.
• Comprehensive experiments were conducted on public datasets and a self-built animal interventional vascular dataset. In addition to quantitative and qualitative reconstruction comparisons, statistical significance analysis, reconstruction-time evaluation, and vascular centerline assessment were included to provide a more rigorous and clinically relevant assessment of the proposed method.

2. RELATED WORK

This section briefly reviews prior literature on classical CT reconstruction, implicit neural representations, and mixture-of-experts models.

2.1. Classical computed tomography reconstruction

Computed tomography reconstruction has long been a fundamental problem in medical imaging. Since the emergence of CT technology in the 1970s, numerous studies have focused on improving reconstruction quality under sparse-view, limited-angle, and noisy acquisition settings. Early algebraic reconstruction techniques, developed for X-ray and electron microscopy, demonstrated advantages in limited-projection scenarios[6]. The simultaneous algebraic reconstruction technique further improved efficiency by updating all pixels in a more coordinated manner and demonstrated superior performance in image quality and computational speed for large-scale reconstruction tasks[3].

To suppress artifacts and preserve structural information under sparse projections, various iterative regularization-based methods have been extensively investigated. For example, total variation regularization has been introduced to improve sparse-view CT quality[7], while total generalized variation has been used to

better preserve edge details and reduce streak artifacts[8]. For limited-angle reconstruction, scale-space anisotropic total variation methods[9] and approaches combining image-gradient regularization with dictionary learning[10] have also been proposed. Although these methods improve reconstruction quality, they typically require iterative optimization, thereby incurring high computational overhead.

2.2. Implicit neural representations

Neural Radiance Fields introduced a powerful way to represent 3D scenes as continuous neural functions and have significantly influenced research in view synthesis and scene reconstruction[5]. Although NeRF was originally designed for red, green, blue (RGB) rendering, its continuous representation framework has inspired many learning-based methods for medical imaging and tomography reconstruction. Several representative extensions have been proposed. Neural adaptive tomography (NeAT) introduced neural adaptive tomography for limited-view reconstruction[11]. TensoRF reduced computational complexity through tensor decomposition while maintaining efficient scene representation[12]. Neural Attenuation Fields adapted the neural field paradigm to sparse-view cone-beam CT by directly modeling attenuation properties[13]. IntraTomo leveraged self-supervised learning for tomography reconstruction through sinogram synthesis and prediction[14]. Other related studies include total-variation-constrained reconstruction for cone-beam CT[15], Mip-NeRF for multi-scale representation[16], voxel-neural hybrid representations[17], and self-supervised masked pretraining for neural field learning[18]. Despite these advances, effectively capturing heterogeneous tissue characteristics and fully exploiting projection-plane information remain open challenges for CT-oriented neural reconstruction.

More recently, NeRF-style and related implicit neural representations have been increasingly investigated for medical image reconstruction. Recent surveys have summarized the broad applications of neural radiance fields in medical imaging, including sparse-view tomography, surgical scene reconstruction, and multimodal 3D representation learning[19]. Structure-aware X-ray NeRF methods further explore incorporating anatomical or projection-domain structural priors for sparse-view X-ray reconstruction[20]. In addition to purely implicit multilayer perceptron (MLP)-based representations, recent studies have introduced explicit or hybrid 3D representations, including 3D Gaussian representations for sparse-view or extremely sparse-view CT/cone-beam computed tomography (CBCT) reconstruction[21,22]. Beyond regularized and meta-learned neural attenuation field methods, researchers have proposed methods to improve the stability of CBCT reconstruction under limited-view conditions[23]. These developments are also related to efficient neural graphics representations such as instant neural graphics primitives, 3D Gaussian splatting, Plenoxels, direct voxel-grid optimization, and unbounded-scene NeRF variants[24-28]. Together, these recent advances indicate that sparse-view medical reconstruction is moving toward stronger priors, hybrid implicit-explicit representations, and more efficient patient-specific optimization.

2.3. Mixture-of-experts models

The mixture-of-experts paradigm was originally introduced to improve model specialization by combining multiple expert subnetworks with a gating mechanism that dynamically selects suitable experts for each input[29]. Later studies extended this idea to deeper neural architectures and showed that expert specialization can improve model capacity and efficiency in complex learning problems[30]. Conditional computation further advanced this direction by activating only the model components relevant to a given input[31]. In large-scale deep learning, sparse expert routing has become a practical strategy for increasing representational power while maintaining manageable computational cost. Representative studies include GShard and Switch Transformer, which demonstrated the effectiveness of sparse gating and expert routing in large neural models[32,33].

In this work, the proposed AME module is inspired by the general mixture of experts (MoE) principle, but it is not intended as a new general-purpose MoE methodology. Instead, we adapt expert specialization and

dynamic routing to sparse-view X-ray attenuation-field reconstruction. Conventional MoE models often use expert routing mainly to increase model capacity under sparse conditional computation. By contrast, AME-Net++ uses expert routing to decompose heterogeneous attenuation modeling into region-adaptive functions. The expert assignment is performed on context-enhanced ray-wise features rather than isolated point-wise features, which better matches the physical characteristics of X-ray projection formation.

Different from large-scale language modeling, where MoE is mainly used to increase model capacity under conditional computation, the MoE design in this work is introduced as a region-adaptive function decomposition mechanism for heterogeneous X-ray attenuation field reconstruction. Medical X-ray reconstruction involves spatially varying attenuation distributions, weak tissue contrast, and region-dependent anatomical structures. Therefore, an adaptive expert routing mechanism can provide a task-specific inductive bias by assigning different local features or attenuation patterns to specialized expert subnetworks. Motivated by these developments, we incorporate a gated mixture-of-experts design into neural attenuation modeling to better address the heterogeneous attenuation properties and structural diversity observed in CT reconstruction tasks.

3. METHOD

AME-Net++ reconstructs a scene-specific three-dimensional representation from sparse multi-view X-ray projections with known acquisition geometry. The central idea is to combine ray-wise contextual modeling with adaptive expert specialization. The former incorporates information from neighboring samples, while the latter applies different learned transformations to heterogeneous attenuation patterns. An information-aware composite loss further emphasizes informative views, local boundaries, and structures at multiple resolutions.

The framework contains four stages. First, individual pixels and contiguous blocks are selected from the projection images, and spatial points are sampled along their corresponding X-ray paths. Second, multi-resolution hash encoding converts the sampled coordinates into spatial features. Third, a sequence module and a gated mixture-of-experts module transform these features into attenuation predictions. Finally, a differentiable X-ray projection model produces synthesized observations, which are compared with the measured projections to optimize the network.

The network is optimized separately for each scene or subject using its available projection observations. The architecture without the proposed composite loss is referred to as AME-Net, whereas the complete framework is termed AME-Net++.

3.1. X-ray imaging model and sampling

The imaging model connects the unknown three-dimensional attenuation distribution to the observed two-dimensional projections. This connection allows projection-domain errors to supervise reconstruction without requiring a reference volume in the optimization objective.

Under an effective monoenergetic approximation and neglecting scattering, X-ray transmission follows the Beer--Lambert law:

$$I(\mathbf{r}) = I_0(\mathbf{r}) \exp\left[-\int_{t_n}^{t_f} \mu(\mathbf{r}(t))\, dt\right]. \tag{1}$$

Here, I_0 represents incident intensity, and the attenuation function describes the local linear attenuation coefficient.

The ray from the X-ray source toward a detector pixel is parameterized as

$$\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}, \qquad \|\mathbf{d}\|_2 = 1. \tag{2}$$

Here, $\mathbf{o}$ is the X-ray source position, $\mathbf{d}$ is the unit direction toward a detector pixel, and t_n and t_f delimit the ray segment inside the reconstruction domain. Because the direction is normalized, the integration parameter corresponds to physical distance.

Equivalently, the logarithmic projection is

$$p(\mathbf{r}) = -\log \frac{I(\mathbf{r})}{I_0(\mathbf{r})} = \int_{t_n}^{t_f} \mu\big(\mathbf{r}(t)\big)\, dt. \tag{3}$$

Equations (1) and (3) distinguish transmitted-intensity observations from line-integral observations. The prediction used for supervision must match the domain of the measured data, and identical fixed preprocessing is applied to the prediction and its corresponding reference.

To approximate the line integral, we uniformly sample N_s points along each ray. The position of the j-th sample is

$$\mathbf{q}_j = \mathbf{o} + t_j\mathbf{d}. \tag{4}$$

The corresponding sampling parameter is

$$t_j = t_n + \left(j - \frac{1}{2}\right)\delta_{\mathbf{r}}, \qquad j = 1, \ldots, N_s. \tag{5}$$

The sampling interval is determined by the ray segment length and the number of samples:

$$\delta_{\mathbf{r}} = \frac{t_f - t_n}{N_s}. \tag{6}$$

We use 320 samples per ray. The sampling interval is constant within a ray but may differ between rays because their intersection lengths with the reconstruction domain vary. Physical integration distances are retained when spatial coordinates are normalized for feature encoding.

Projection-domain sampling combines individual pixels and contiguous image blocks. Pixel-level sampling provides coverage of the observations, whereas block-level sampling preserves the spatial arrangement required for perceptual and multi-scale supervision. Each optimization step uses 2,048 rays, including 1,024 block-level samples and 1,024 pixel-level samples. A threshold-based foreground selection strategy guides sampling toward structurally informative regions.

3.2. Multi-resolution ray-wise feature encoding

Spatial encoding provides a learnable description of each sampled location. Using multiple resolutions enables the representation to capture both broad anatomical variation and fine spatial detail without relying on a large coordinate-only MLP.

Each sampled point is normalized to the encoding domain and mapped to a C-dimensional feature vector using multi-resolution hash encoding[24]. The encoded features are stacked in their order along the ray:

$$\mathbf{X} = \begin{bmatrix} \gamma\big(\mathcal{T}(\mathbf{q}_1)\big)^\top \\ \vdots \\ \gamma\big(\mathcal{T}(\mathbf{q}_{N_s})\big)^\top \end{bmatrix} \in \mathbb{R}^{N_s \times C}. \tag{7}$$

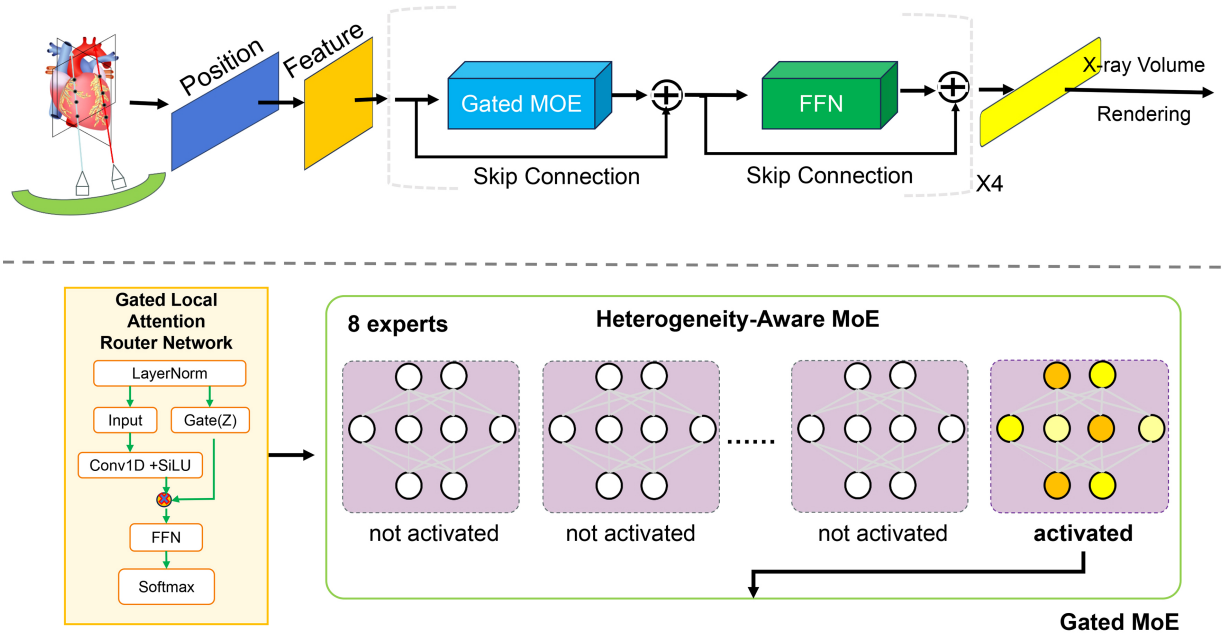


Figure 1. Architecture of the proposed gated mixture-of-experts module for heterogeneous attenuation modeling and adaptive reconstruction. MoE: Mixture of experts; FFN: feed-forward network; SiLU: sigmoid linear unit.

Here, the fixed transform normalizes coordinates to the encoding domain, and the hash encoder maps each normalized position to its spatial feature vector. The matrix $\mathbf{X}$ therefore contains one feature row per sampled point.

At this stage, each row describes an individual spatial location. The subsequent sequence module incorporates neighboring information along the ray, after which the expert module performs adaptive feature transformation.

3.3. Context-enhanced mixture-of-experts reconstruction

The reconstruction network addresses two complementary requirements. The sequence module enriches each sample feature with local context, while the mixture-of-experts module adapts the feature transformation to heterogeneous attenuation patterns. Their combination reduces the need for a single shared transformation to represent all anatomical variations.

The network architecture is illustrated in Figure 1. To control computation and memory usage, the ray-wise feature sequence is divided into K non-overlapping segments:

$$\{\mathbf{X}^{(k)}\}_{k=1}^{K} = \mathrm{Partition}(\mathbf{X}, K), \qquad \mathbf{X}^{(k)} \in \mathbb{R}^{L\times C}. \tag{8}$$

The length of each segment is

$$L = \frac{N_s}{K}. \tag{9}$$

The segment count is chosen to divide the number of ray samples exactly.

Each segment is first processed by the sequence module with a residual connection:

$$\mathbf{H}^{(k)} = \mathbf{X}^{(k)} + \mathrm{SeqProcess}\left(\mathbf{X}^{(k)}\right). \tag{10}$$

The context-enhanced features are then processed by the mixture-of-experts module with another residual connection:

$$\mathbf{Y}^{(k)} = \mathbf{H}^{(k)} + \mathrm{MoE}\left(\mathbf{H}^{(k)}\right). \tag{11}$$

The residual paths retain the original features while allowing the two modules to learn complementary corrections. The processed segments are subsequently restored to their original depth order.

3.3.1. Sequence processing module

An isolated spatial feature does not explicitly describe its neighboring samples. The sequence module addresses this limitation by aggregating local information along the ray. Consequently, the router and expert subnetworks receive context-enhanced features rather than isolated point features.

For a generic segment, layer normalization and a linear projection produce a feature branch $\mathbf{A}$ and a modulation branch $\mathbf{Z}$:

$$[\mathbf{A}, \mathbf{Z}] = \mathrm{Split}\left[\mathrm{LayerNorm}(\mathbf{X}^{(k)})\mathbf{W}_{\mathrm{proj}}\right]. \tag{12}$$

The projection matrix has C input channels and $2C'$ output channels, so both branches contain L samples with C' channels.

A depth-wise one-dimensional convolution aggregates neighboring samples in the feature branch:

$$\mathbf{A}_{\mathrm{seq}} = \mathrm{SiLU}\left[\mathrm{DWConv1D}_{\kappa}(\mathbf{A})\right]\mathbf{W}_{\mathrm{seq}}. \tag{13}$$

The modulation branch controls the contribution of the aggregated features, followed by an output projection:

$$\mathrm{SeqProcess}(\mathbf{X}^{(k)}) = \left[\mathbf{A}_{\mathrm{seq}} \odot \mathrm{SiLU}(\mathbf{Z})\right]\mathbf{W}_{\mathrm{out}}. \tag{14}$$

The convolution operates along the sample dimension with length-preserving padding. The sequence projection preserves C' channels, and the output projection maps the features back to C channels, ensuring compatibility with the residual connection. The symbol between the two feature branches denotes element-wise multiplication.

This module captures local ray-wise context, with its receptive field determined by the convolution kernel and network depth. Independent segment processing does not introduce cross-segment communication. The contextual operation is an architectural prior for the inverse problem, rather than a claim that the physical attenuation coefficient depends on preceding ray samples.

3.3.2. Gated mixture-of-experts module

After contextual processing, the mixture-of-experts module applies different learned transformations to different sample features. This provides a flexible alternative to requiring one shared MLP to model all local attenuation patterns.

Gated routing. For a generic segment, the context-enhanced features are first normalized:

$$\overline{\mathbf{H}} = \mathrm{LayerNorm}(\mathbf{H}^{(k)}). \tag{15}$$

The normalized features are then mapped to expert probabilities:

$$\mathbf{G} = \mathrm{Softmax}_E\left(\overline{\mathbf{H}}\mathbf{W}_g + \mathbf{1}_L\mathbf{b}_g^\top\right). \tag{16}$$

The router maps C-channel features to probabilities over E experts, using a learnable weight matrix and bias. Softmax is evaluated across the expert dimension. We use eight experts in the implementation.

For token n, Top-1 routing selects the expert with the largest probability:

$$e_n^* = \arg\max_{e\in\{1,\ldots,E\}} G_{n,e}. \tag{17}$$

The selected index determines which expert processes the token. Routing is learned from reconstruction supervision and does not require manually specified tissue labels or anatomical partitions.

Expert processing. Let $\mathbf{h}_n^\top$ denote the n-th row of the normalized feature matrix. Each expert is a three-layer MLP with Gaussian error linear unit (GELU) activations between successive linear layers.

The first hidden representation is computed as

$$\mathbf{a}_{e,n}^{(1)} = \mathrm{GELU}\left[\left(\mathbf{W}_e^{(1)}\right)^\top \mathbf{h}_n + \mathbf{b}_e^{(1)}\right]. \tag{18}$$

The second hidden representation is

$$\mathbf{a}_{e,n}^{(2)} = \mathrm{GELU}\left[\left(\mathbf{W}_e^{(2)}\right)^\top \mathbf{a}_{e,n}^{(1)} + \mathbf{b}_e^{(2)}\right]. \tag{19}$$

The final expert output is

$$\mathbf{F}_e(\mathbf{h}_n) = \left(\mathbf{W}_e^{(3)}\right)^\top \mathbf{a}_{e,n}^{(2)} + \mathbf{b}_e^{(3)}. \tag{20}$$

The first layer maps C channels to the hidden width C_h, the second preserves that width, and the final layer returns C output channels. Bias dimensions match the corresponding layer outputs.

Only the selected expert is evaluated for each token. Its output is weighted by the selected routing probability:

$$\mathbf{v}_n = G_{n,e_n^*}\mathbf{F}_{e_n^*}(\mathbf{h}_n). \tag{21}$$

The weighted token outputs are stacked to form the mixture-of-experts output:

$$\mathrm{MoE}(\mathbf{H}^{(k)}) = \begin{bmatrix}\mathbf{v}_1^\top \\ \vdots \\ \mathbf{v}_L^\top\end{bmatrix}. \tag{22}$$

The processed tokens are restored to their original order before the residual addition. Retaining the selected soft probability provides a gradient path to the router, although the discrete expert index itself is not differentiated.

Through this mechanism, different experts can learn complementary transformations for different feature patterns. Top-1 routing reduces expert computation relative to evaluating every expert for every token, but does not by itself guarantee balanced utilization or a one-to-one correspondence between experts and tissue categories.

3.3.3. Attenuation prediction and projection rendering

After expert processing, the segments are concatenated in their original depth order:

$$\mathbf{Y} = \mathrm{Concat}_{\mathrm{depth}}\left(\mathbf{Y}^{(1)}, \ldots, \mathbf{Y}^{(K)}\right). \tag{23}$$

A shared prediction head then converts each feature into a positive attenuation estimate:

$$\widehat{\mu}_j(\mathbf{r}) = \mathrm{Softplus}\left(\mathbf{w}_{\mu}^{\top}\mathbf{y}_j + b_{\mu}\right). \tag{24}$$

Here, $\mathbf{y}_j^{\top}$ is the j-th row of the merged feature matrix, and the prediction head maps its C channels to a scalar. Softplus produces positive values that can approach zero.

The sample predictions are integrated to obtain a logarithmic projection:

$$p_{\mathrm{pred}}(\mathbf{r}) = \delta_{\mathbf{r}} \sum_{j=1}^{N_s} \widehat{\mu}_j(\mathbf{r}). \tag{25}$$

When transmitted-intensity observations are required, the logarithmic projection is converted into intensity:

$$I_{\mathrm{pred}}(\mathbf{r}) = I_0(\mathbf{r}) \exp[-p_{\mathrm{pred}}(\mathbf{r})]. \tag{26}$$

This differentiable rendering step connects the three-dimensional sample predictions to the measured projection data. Projection errors can therefore update all learnable components of the reconstruction network.

Because the network uses ray-wise context, its sample predictions are explicitly written as ray-conditioned estimates. The architecture does not inherently guarantee identical predictions for the same spatial location queried under different ray contexts.

3.4. Information-aware multi-scale loss

Projection supervision must address three complementary requirements: fidelity across views, preservation of local boundaries, and consistency across spatial scales. We therefore combine an entropy-weighted projection loss, a local gradient-aware loss, and a multi-scale auxiliary loss.

For compact notation, $Q_{\mathrm{pred}}^{(i)}$ and $Q_{\mathrm{gt}}^{(i)}$ denote the predicted and observed projections of view i in the selected supervision domain after identical fixed preprocessing. Reference values used for histogram construction are normalized to the interval from zero to one.

Let N_v be the number of views in a batch and R_i the number of sampled rays from view i. The residual at a detector coordinate is

$$e_i(\mathbf{u}) = Q_{\text{pred}}^{(i)}(\mathbf{u}) - Q_{\text{gt}}^{(i)}(\mathbf{u}). \tag{27}$$

The basic fidelity penalty combines squared and absolute errors:

$$\rho(e) = \lambda_{\text{MSE}} e^2 + \lambda_{\text{L1}} |e|. \tag{28}$$

The nonnegative coefficients balance the two error terms. Pixel-wise losses are averaged over the evaluated ray samples rather than the number of pixels in an unrendered full projection.

3.4.1. Entropy-weighted projection loss

Different viewing angles produce different projection intensity distributions. Entropy weighting uses this variation as a heuristic for assigning different importance to the observed views.

The entropy of each reference projection is computed as

$$\mathcal{E}_i = - \sum_{\substack{b=1 \\ P_i(b)>0}}^{B} P_i(b) \log P_i(b). \tag{29}$$

Here, $P_i(b)$ is the fraction of reference pixels falling into histogram bin b. Before histogram construction, each reference projection is represented in the normalized intensity range from 0 to 1. The histogram contains 256 uniformly spaced bins; therefore, B is set to 256 in all experiments. Histogram bins with zero probability are excluded before evaluating the logarithm. Consequently, no stabilization constant is added inside the logarithm, and empty bins contribute zero to the entropy, exactly as specified in Equation (29). The entropy is computed only from the observed reference projections and does not receive gradients.

The normalized batch weight of each view is

$$w_i^{\text{ent}} = \frac{\exp(\mathcal{E}_i)}{\sum_{a=1}^{N_v} \exp(\mathcal{E}_a)}. \tag{30}$$

When only one view participates in the weighted objective, its weight equals one.

The entropy-weighted projection loss is

$$\mathcal{L}_{\text{ent}} = \sum_{i=1}^{N_v} w_i^{\text{ent}} \left[\frac{1}{R_i} \sum_{m=1}^{R_i} \rho\big(e_i(\mathbf{u}_{i,m})\big) + \lambda_{\text{LPIPS}} \ell_i^{\text{perc}} \right]. \tag{31}$$

The sampled detector locations are indexed by m. The pixel-wise terms constrain numerical agreement, while the perceptual term provides complementary patch-level supervision.

Specifically, the perceptual term averages learned perceptual image patch similarity (LPIPS) over the contiguous blocks sampled from each view:

$$\ell_i^{\text{perc}} = \frac{1}{J_i} \sum_{b=1}^{J_i} \text{LPIPS}\left(\mathcal{T}_{\text{perc}}(\mathcal{P}_{\text{pred}}^{(i,b)}), \mathcal{T}_{\text{perc}}(\mathcal{P}_{\text{gt}}^{(i,b)}) \right). \tag{32}$$

Here, J_i is the number of valid contiguous blocks sampled from view i. The perceptual loss is implemented using the official PyTorch LPIPS package of Zhang *et al.*[34], with the learned LPIPS model and the pretrained AlexNet backbone. The model is instantiated using lpips.LPIPS(net='alex', lpips=True). For each predicted

and reference projection patch, the single-channel floating-point input is first represented in the normalized range from 0 to 1. The channel is then replicated three times to form an input tensor with dimensions $N \times 3 \times H \times W$. The resulting tensor is linearly mapped from the interval [0, 1] to the interval [-1, 1], as required by the official LPIPS implementation. Identical preprocessing is applied to the predicted and reference patches. The parameters of the pretrained AlexNet backbone and the learned LPIPS linear calibration layers remain fixed during reconstruction optimization, although gradients with respect to the predicted input patches are retained. LPIPS is computed separately for each valid contiguous projection block and is subsequently averaged over the participating blocks and views. No LPIPS loss is applied to independently sampled pixels because these samples do not form spatially contiguous image patches.

Histogram entropy measures intensity diversity rather than directly providing anatomical information. Its role is therefore to provide heuristic view weighting, not a direct anatomical information measure.

3.4.2. Local gradient-aware loss

View-level weighting cannot distinguish an anatomical boundary from an uninformative region within the same projection. The local loss addresses this limitation by assigning higher weights to pixels near strong reference-image gradients.

Horizontal and vertical Sobel operators produce a gradient magnitude map:

$$g_i(\mathbf{u}) = \sqrt{\left[\nabla_u Q_{\mathrm{gt}}^{(i)}(\mathbf{u})\right]^2 + \left[\nabla_v Q_{\mathrm{gt}}^{(i)}(\mathbf{u})\right]^2}. \tag{33}$$

The gradient magnitude is normalized by its maximum value:

$$D_i(\mathbf{u}) = \frac{g_i(\mathbf{u})}{\max_{\mathbf{v}} g_i(\mathbf{v}) + \epsilon_D}. \tag{34}$$

The positive stabilization constant prevents division by zero for constant projections.

The normalized map is then spatially smoothed:

$$\widetilde{D}_i = \mathrm{AvgPool}_{3\times 3}(D_i). \tag{35}$$

Average pooling preserves the map dimensions and reduces sensitivity to isolated high-frequency fluctuations. Both normalization and smoothing are applied to reference data, so the resulting weights remain fixed during optimization.

The local loss is evaluated at the sampled ray locations:

$$\mathcal{L}_{\mathrm{local}} = \frac{1}{N_v}\sum_{i=1}^{N_v}\frac{1}{R_i}\sum_{m=1}^{R_i}\left[1 + \widetilde{D}_i(\mathbf{u}_{i,m})\right]\rho\left(e_i(\mathbf{u}_{i,m})\right). \tag{36}$$

The unit baseline weight retains supervision in low-gradient regions, while the additional gradient weight emphasizes boundaries and other edge-rich structures. LPIPS is not included in this term because the weighting is defined at individual sampled pixels.

3.4.3. Multi-scale auxiliary loss

The multi-scale term complements original-resolution supervision by emphasizing structures that remain visible after downsampling. It encourages agreement in broader spatial patterns while the preceding terms retain fine-scale fidelity.

Since independently sampled pixels do not form a dense image, multi-scale supervision is applied only to contiguous block samples.

The predicted patches are downsampled by average pooling:

$$\mathcal{P}_{\mathrm{pred},s}^{(i,b)} = \mathrm{AvgPool}_{s\times s}\left(\mathcal{P}_{\mathrm{pred}}^{(i,b)}\right), \qquad s \in \{2,4\}. \tag{37}$$

The corresponding reference patches are downsampled using the same operation:

$$\mathcal{P}_{\mathrm{gt},s}^{(i,b)} = \mathrm{AvgPool}_{s\times s}\left(\mathcal{P}_{\mathrm{gt}}^{(i,b)}\right), \qquad s \in \{2,4\}. \tag{38}$$

Average pooling uses a stride equal to the scale factor. Patch dimensions are selected to support both downsampling factors, and the resulting patches must remain compatible with the perceptual backbone.

At each scale, the entropy-weighted and local losses follow the definitions above, with pixel averages evaluated over the downsampled patches. View entropy is recomputed from the corresponding downsampled reference projection, while local gradient maps are recomputed from the downsampled reference patches.

The auxiliary objective is

$$\mathcal{L}_{\mathrm{ms}} = \sum_{s\in\{2,4\}} \alpha_s \left(\mathcal{L}_{\mathrm{ent},s} + \mathcal{L}_{\mathrm{local},s}\right), \qquad \alpha_s \geq 0. \tag{39}$$

The scale coefficients control the relative contribution of each reduced resolution. Original-resolution supervision is excluded from this auxiliary term because it is already provided by the first two losses.

3.4.4. Overall objective

The complete objective combines the three complementary forms of supervision:

$$\mathcal{L}_{\mathrm{total}} = \beta_{\mathrm{ent}}\mathcal{L}_{\mathrm{ent}} + \beta_{\mathrm{local}}\mathcal{L}_{\mathrm{local}} + \beta_{\mathrm{ms}}\mathcal{L}_{\mathrm{ms}}. \tag{40}$$

The entropy, local, and multi-scale coefficients are set to 0.8, 0.4, and 0.3, respectively. These coefficients were manually predefined before the full experiments and were subsequently kept fixed for all datasets, anatomical scenes, and compared experimental settings. They were not selected using a particular scene or subject, and no scene-specific tuning or systematic hyperparameter search was performed. The coefficients balance supervision across views, locations, and resolutions and are not required to sum to one.

The global and local terms intentionally share fidelity penalties with different weighting rules. Their contributions are complementary but not mathematically independent. Detailed sampling settings, encoder dimensions, and remaining loss hyperparameters are reported with the implementation settings.

3.5. Optimization and reconstruction outputs

All learnable components are jointly optimized using Adam with an initial learning rate of 0.0001. In each iteration, selected projection samples are converted into calibrated rays, processed by the reconstruction network, and rendered into predicted observations. The composite projection loss in Equation (40) is then backpropagated through the entire pipeline. A paired reference volume is not required by this objective.

After optimization, projections at additional viewing angles are synthesized using the same ray construction and rendering procedure. For quantitative novel-view evaluation, the evaluated projection angles are excluded from the optimization observations. This assesses within-scene synthesis rather than cross-subject prediction by a pretrained model.

For volumetric reconstruction, AME-Net++ explicitly extracts the learned three-dimensional attenuation field by querying the optimized neural representation on a dense regular grid. During evaluation, an axis-aligned grid is generated according to the tomographic iterative GPU-based reconstruction (TIGRE) cone-beam geometry, with nVoxel, dVoxel, and offOrigin converted to meters. The grid resolution is matched to the reference CT volume, which is set to 128^3 by default, and the physical voxel spacing is preserved to maintain the correct spatial scale. Unlike ray sampling during training, the grid extraction stage is deterministic and does not introduce stochastic perturbations. The grid coordinates are flattened into an $N \times 3$ coordinate matrix and processed in chunks of N_{chunk} =409,600 points for memory-efficient inference. Each coordinate is first encoded by the multi-resolution hash encoder. The encoded features are then fed into the proposed AME reconstruction module, where local contextual features are modeled by the sequence-processing branch and heterogeneous attenuation patterns are further represented by the gated mixture-of-experts branch. The learnable router selects the most suitable expert for each feature token, allowing different spatial regions and attenuation patterns to be reconstructed by specialized expert subnetworks. Finally, a shared attenuation prediction head maps the expert-enhanced features to positive attenuation coefficients. The predicted coefficients from all chunks are concatenated and reshaped into the original three-dimensional volume, yielding a single attenuation value for each voxel. Therefore, the reconstructed volume is obtained by direct neural field extraction from the proposed adaptive mixture-of-experts representation. This extraction procedure does not rely on analytical reconstruction, iterative voxel updates, or external post-processing. Each voxel is evaluated once at its grid-center coordinate. No multi-directional sampling, sliding-window inference, test-time augmentation, or ensemble averaging is used. The network parameters remain fixed after optimization, and the same attenuation head used for differentiable projection rendering is used for volumetric extraction, which keeps the 2D projection synthesis and 3D reconstruction results consistent.

4. EXPERIMENTS

A systematic evaluation of the proposed method was conducted using public CT datasets covering multiple anatomical regions, and comparisons were performed against representative reconstruction baselines. The effectiveness of the proposed network architecture and loss function design is demonstrated through both quantitative and qualitative experiments.

To provide a more comprehensive and rigorous evaluation, we further supplemented the experiments with statistical significance analysis, reconstruction-time comparison, and additional validation on an in-house animal interventional vascular dataset. This dataset was collected during interventional experiments at an animal experimental facility under a study protocol approved by the Ethics Committee of Shanghai Xinova Medical Research Co., Ltd. (approval No. XNM-YX-20221103-01). The collection and use of the dataset were covered by this approval. These supplementary experiments assess not only reconstruction accuracy but also computational efficiency, robustness, and potential clinical applicability.

4.1. Experimental settings

To simulate sparse-view reconstruction scenarios, we used several publicly available CT datasets covering the head, foot, chest, pancreas, abdomen, leg, and aneurysm regions. These datasets were obtained from LIDC-IDRI[35], MIDA[36], DeepOrgan[37], and VOLVIS[38]. Public CT datasets are mainly used to provide a broader validation of AME-Net++ across different anatomical structures and imaging scenarios.

For the public CT datasets, projection images were generated using the TIGRE tomography toolbox under a cone-beam X-ray geometry, following the standard protocol adopted in related neural attenuation-field reconstruction studies. For each scene, 100 projection views were uniformly generated over a full 360° rotation, with an angular interval of 3.6° between consecutive views. The projections were divided into two interleaved subsets according to their angular order. Fifty views with even indices were used for scene-specific optimization, while the remaining 50 views with odd indices were held out exclusively for novel-view projection evaluation. In addition, 3% synthetic noise was added to the generated projections to simulate realistic imaging degradation.

We further constructed an in-house interventional vascular dataset collected at an animal experimental facility under the approval of the Ethics Committee (approval No. XNM-YX-20221103-01), to evaluate the feasibility of vascular reconstruction in a clinically relevant interventional setting. The preoperative CTA data were acquired using a 256-slice CT scanner. The vascular structures were manually extracted from the CTA images by experienced clinicians and used as the reference 3D vascular models. During the interventional experiments, angiographic images were acquired from 10 different projection views and then used for vascular reconstruction.

The 50-view setting is mainly used as a controlled sparse-view benchmark for complex volumetric reconstruction involving multiple anatomical structures, including organs, vessels, bones, and surrounding tissues. In contrast, the 10-view limited-angle setting in the animal interventional vascular dataset is used for clinically oriented vascular reconstruction, where the reconstruction target is more focused and mainly involves contrast-enhanced vascular structures.

4.2. Implementation details

AME-Net++ was implemented in PyTorch and optimized using the Adam optimizer for 2,000 epochs with an initial learning rate of 1e-4. The X-ray batch size was set to 2,048 rays, consisting of 1,024 rays from block-level sampling and 1,024 rays from pixel-level sampling. Uniform ray sampling was performed with 320 points per ray. For each public CT scene, 100 projection views are uniformly generated over a full 360° acquisition trajectory. The 50 views with even zero-based indices are used for optimization, corresponding to the angular sequence.

$$\{0^\circ, 7.2^\circ, 14.4^\circ, \ldots, 352.8^\circ\}.$$

The remaining 50 views with odd zero-based indices are held out exclusively for novel-view evaluation, corresponding to

$$\{3.6^\circ, 10.8^\circ, 18.0^\circ, \ldots, 356.4^\circ\}.$$

Thus, every held-out projection is located midway between two adjacent optimization views. None of the held-out projections are used for network optimization, loss computation, hyperparameter determination, or early stopping. The reported novel-view peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) values are calculated exclusively from these 50 held-out projections. For the animal interventional vascular reconstruction experiment, 10 intraoperative angiographic projection views are used for each case. The number of experts in the gated mixture-of-experts module is set to 8, and the loss weights β_{ent}, β_{local}, and β_{ms} are set to 0.8, 0.4, and 0.3, respectively. All experiments were conducted on an NVIDIA A100 GPU. PSNR and structural similarity index are used as the main quantitative evaluation metrics.

It should be clarified that AME-Net++ follows a NeRF-style implicit reconstruction paradigm and does not use a conventional fixed training/testing split. For each anatomical scene or patient, the framework establishes a mapping relationship between the available projection observations and the corresponding 3D spatial representation. Since each patient has a different anatomical space and vascular morphology, a new mapping relationship needs to be established for each patient based on the projection images acquired from that patient. The resulting 3D reconstruction reflects the capability of the proposed implicit representation and network design to recover patient-specific vascular structures from sparse-view observations.

Therefore, this work evaluates the quality of the reconstructed 3D structure under sparse-view conditions, rather than cross-patient generalization in the conventional supervised-learning sense. The same network design, loss functions, and optimization settings are applied across cases to ensure a fair evaluation of the proposed reconstruction framework.

Quantitative metric implementation. For both novel-view projection synthesis and volumetric reconstruction, the predicted and reference data were evaluated in the same normalized intensity domain. All data were represented as single-channel floating-point arrays and normalized to the range [0, 1] using the fixed transformation adopted during data preprocessing. The same transformation was applied to each corresponding prediction–reference pair. No independent per-image, per-volume, or per-method normalization was performed. Values outside the valid range were clipped to [0, 1] before evaluation, and no histogram matching was applied. PSNR and SSIM were calculated using the peak_signal_noise_ratio and structural_similarity functions from skimage.metrics, respectively, with data_range=1.0. For novel-view projection synthesis, PSNR and SSIM were calculated separately for each of the 50 held-out projections and then averaged over all held-out views. For volumetric reconstruction, the metrics were calculated directly over the complete reconstructed and reference volumes at the same voxel resolution. The same evaluation procedure was applied to all compared methods.

4.3. Self-built animal interventional vascular dataset

To further validate the proposed method under real interventional imaging conditions, an in-house animal interventional vascular dataset was collected under a protocol approved by the Animal Ethics Committee of Shanghai Xinova Medical Research Co., Ltd. on November 3, 2022 (approval No. XNM-YX-20221103-01). Following ethical approval, 31 Bama miniature pigs (16 females and 15 males), aged 12-14 months and weighing 30-45 kg at the time of the experiments, were purchased on November 4, 2022. The laboratory animal production license associated with the animal source was No. SCXK (Su) 2021-0007. The experiments were conducted at the Laboratory Animal Center of Shanghai Xinova Medical Research Co., Ltd. from November 13 to 20, 2022. Anesthesia was induced by subcutaneous injection of ketamine hydrochloride and maintained with inhaled isoflurane in oxygen following endotracheal intubation. Electrocardiographic activity and oxygen saturation were continuously monitored throughout the procedures. After anesthesia, the animals were positioned supine, and femoral venous access was established using the Seldinger technique, followed by placement of a 6-F vascular sheath. Preoperative contrast-enhanced CT angiography of the thoracoabdominal aorta was performed using a 256-slice CT scanner, with a total contrast volume of 70 mL administered at an injection rate of 3 mL/s. Vascular structures were manually extracted from the CTA images by experienced clinicians to serve as the reference 3D vascular models. During navigation functionality validation, vascular projection images were acquired from 10 different angles and used as input for vascular reconstruction to evaluate the method's ability to assist surgical procedures. During the experiment, one animal developed myocardial ischemia, and its data were excluded from the final analysis. Data from the remaining 30 animals were therefore included. Completion of the navigation functionality validation marked the planned experimental endpoint, after which the surviving animals were euthanized.

5. RESULTS

We compare AME-Net++ with several representative baselines, including FDK, IntraTomo, NeRF, TensoRF, and neural attenuation fields (NAF), on both 3D reconstruction and projection synthesis tasks. As shown in Tables 1 and 2, our framework achieves the best or near-best results on most anatomical scenes, demonstrating strong effectiveness and robustness under sparse-view conditions. In addition to the mean quantitative metrics, Table 3 presents the statistical significance analysis. Table 4 reports the ablation study results for AME-Net++, while Table 5 summarizes the validation results on our self-built animal interventional vascular dataset. Finally, Table 6 compares the reconstruction times of the evaluated methods.

It should be emphasized that the evaluation follows a projection-based implicit reconstruction protocol. For each scene or patient, the available projection observations are used to establish the mapping relationship to the corresponding 3D spatial representation. Therefore, the reported results assess the reconstruction capability of the proposed representation, network architecture, loss function, and optimization strategy under sparse-view conditions, rather than cross-patient prediction performance in the conventional supervised-learning sense.

3D reconstruction. Table 1 reports the quantitative 3D reconstruction results on seven CT scenes. On average, AME-Net achieves the best overall performance among all compared methods, reaching a PSNR of 35.06 dB and an SSIM of 0.9688. By further incorporating the proposed composite loss, AME-Net++ improves the average results to 35.71 and 0.9732, respectively. In contrast, the strongest baseline, NAF, reaches 33.43 in PSNR and 0.9446 in SSIM. These results highlight clear gains in both reconstruction accuracy and structural similarity.

The proposed framework demonstrates consistent advantages across anatomically diverse regions. On the chest scene, AME-Net++ achieves 34.01 in PSNR and 0.9686 in SSIM, surpassing NAF (33.05/0.9581). On the abdomen scene, it further improved the performance to 35.44 dB and 0.9635, ranking first among all compared methods. The gains become even more evident on structurally complex scenes such as aneurysm, head, and leg, where AME-Net++ reaches 41.11/0.9970, 39.67/0.9897, and 44.62/0.9983, respectively. These results indicate that AME-Net++ captures heterogeneous attenuation patterns more effectively and preserves fine anatomical structures more faithfully than previous methods.

The reported mean values summarize the numerical performance differences, while paired two-sided Wilcoxon signed-rank tests provide complementary statistical evidence for the performance comparisons across the seven anatomical scenes. As reported in Table 3, AME-Net++ achieves nominally significant improvements over AME-Net in 3D reconstruction SSIM ($P = 0.031$) and projection synthesis PSNR ($P = 0.016$) at the uncorrected significance threshold of 0.05. In contrast, the improvements in 3D reconstruction PSNR ($P = 0.063$) and projection synthesis SSIM ($P = 0.219$) do not reach this threshold and should therefore be interpreted as numerical performance trends rather than statistically significant gains.

All reported P-values are uncorrected, and no multiple-comparison correction was applied. Paired observations with zero differences were excluded before ranking, reducing the effective sample size for some comparisons. The statistical results suggest that the observed differences generally favor AME-Net++, but this overall tendency alone does not establish statistical significance. Given the limited number of scenes and the multiple comparisons, these findings should be interpreted as exploratory rather than confirmatory.

Projection synthesis. Table 2 summarizes the quantitative results on novel-view projection synthesis across multiple medical imaging scenes. AME-Net again shows a substantial advantage over all competing methods, achieving an average PSNR/SSIM of 50.88/0.9993. With the proposed composite loss, AME-Net++ further

Table 1. Comparison of different methods on CT reconstruction datasets

Scene	FDK		IntraTomo		NeRF		TensoRF		NAF		AME-Net		AME-Net++	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Chest	22.89	0.7861	22.04	0.7460	28.40	0.8925	30.13	0.9308	33.05	0.9581	33.71	0.9681	34.01	0.9686
Abdomen	22.63	0.6030	28.05	0.8754	29.71	0.9049	31.51	0.9073	34.45	0.9501	34.21	0.9530	35.44	0.9635
Aneurysm	28.07	0.7295	30.32	0.9652	31.97	0.9353	37.36	0.9871	37.73	0.9871	41.11	0.9958	41.11	0.9970
Foot	24.53	0.6000	31.43	0.9127	30.03	0.9072	30.46	0.9153	31.63	0.9363	32.13	0.9292	32.34	0.9414
Head	26.17	0.7155	31.07	0.9303	34.15	0.9672	35.53	0.9672	36.46	0.9743	39.67	0.9897	39.67	0.9897
Leg	28.48	0.6690	36.41	0.9882	39.27	0.9938	40.70	0.9923	41.28	0.9940	41.89	0.9963	44.62	0.9983
Pancreas	9.64	0.1232	16.01	0.7865	17.73	0.8614	19.75	0.7737	19.41	0.8126	22.70	0.9494	22.77	0.9536
Average	23.20	0.6038	27.90	0.8863	30.18	0.9232	32.21	0.9248	33.43	0.9446	35.06	0.9688	**35.71**	**0.9732**

The best average PSNR and SSIM values are highlighted in bold. CT: Computed tomography; FDK: Feldkamp–Davis–Kress; NeRF: neural radiance fields; NAF: neural attenuation fields; PSNR: peak signal-to-noise ratio; SSIM: structural similarity index measure.

Table 2. Comparison of projection synthesis performance on multiple medical imaging scenes

Scene	IntraTomo		NeRF		TensoRF		NAF		AME-Net		AME-Net++	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Chest	28.95	0.9915	36.16	0.9988	23.61	0.9402	42.37	0.9993	49.26	0.9994	50.35	0.9995
Leg	42.53	0.9976	45.92	0.9989	41.40	0.9969	50.87	0.9988	53.18	0.9995	59.46	0.9998
Foot	39.48	0.9979	41.05	0.9989	37.73	0.9929	38.35	0.9913	46.91	0.9994	47.43	0.9994
Head	34.83	0.9977	29.76	0.9921	34.43	0.9878	30.17	0.9531	53.69	0.9996	55.69	0.9997
Pancreas	20.03	0.8537	19.85	0.8560	29.24	0.8031	36.23	0.8844	49.87	0.9978	50.19	0.9981
Abdomen	27.64	0.9646	24.62	0.9559	27.38	0.8730	37.59	0.9855	51.28	0.9995	52.25	0.9996
Aneurysm	20.81	0.9621	24.97	0.9792	47.99	0.9997	39.62	0.9990	52.00	0.9998	54.76	0.9996
Average	30.61	0.9664	31.76	0.9685	34.54	0.9419	39.31	0.9731	50.88	0.9993	**52.88**	**0.9994**

The best average PSNR and SSIM values are highlighted in bold. NeRF: Neural radiance fields; NAF: neural attenuation fields; PSNR: peak signal-to-noise ratio; SSIM: structural similarity index measure.

improves the average PSNR to 52.88 while maintaining strong synthesis quality overall. Compared with NAF (39.31), the proposed framework shows a much stronger capability for novel-view projection generation.

Table 3. Statistical comparison of AME-Net++ with the remaining competing methods over the seven anatomical scenes

Compared method	PSNR			SSIM		
	Effective *n*	*P*-value	RBC	Effective *n*	*P*-value	RBC
CT reconstruction						
FDK	7	0.016	1.000	7	0.016	1.000
IntraTomo	7	0.016	1.000	7	0.016	1.000
NeRF	7	0.016	1.000	7	0.016	1.000
TensoRF	7	0.016	1.000	7	0.016	1.000
NAF	7	0.016	1.000	7	0.016	1.000
AME-Net	5	0.063	1.000	6	0.031	1.000
Projection synthesis						
IntraTomo	7	0.016	1.000	7	0.016	1.000
NeRF	7	0.016	1.000	7	0.016	1.000
TensoRF	7	0.016	1.000	7	0.031	0.929
NAF	7	0.016	1.000	7	0.016	1.000
AME-Net	5	0.063	1.000	6	0.025	0.619

P-values were obtained from separate two-sided Wilcoxon signed-rank tests without correction for multiple comparisons. Zero paired differences were excluded before ranking, and tied absolute differences were assigned average ranks. The effective sample size reported in the table is the number of nonzero paired differences included in each test. Effect sizes are reported as RBC. Differences were defined as AME-Net++ minus the comparison method; positive RBC values therefore indicate better performance of AME-Net++. Given the limited number of scenes and the multiple pairwise comparisons, the *P*-values should be interpreted as exploratory rather than confirmatory statistical evidence. PSNR: Peak signal-to-noise ratio; SSIM: structural similarity index measure; RBC: rank-biserial correlation; CT: computed tomography; FDK: Feldkamp–Davis–Kress; NeRF: neural radiance fields; NAF: neural attenuation fields.

Table 4. Quantitative ablation study of the proposed AME-Net++ components on the novel-view projection synthesis task

MoE module	Sequence module	Entropy loss	local gradient-aware loss	Multi-scale loss	PSNR (dB)	SSIM
					44.97	0.9985
√					48.73	0.9990
√	√				50.88	0.9993
√	√	√			51.57	0.9993
√	√	√	√		51.89	0.9993
√	√	√	√	√	**52.88**	**0.9994**

For the novel-view projection synthesis task, we start with a baseline without MoE or sequence modules and progressively integrate the architectural modules to form AME-Net. We then sequentially incorporate the three joint loss formulations to obtain the optimal AME-Net++ configuration. Bold indicates the best performance. PSNR: Peak signal-to-noise ratio; SSIM: structural similarity index measure; MoE: mixture of experts.

Table 5. Comparison results on our self-built animal interventional vascular dataset

Method	Multiview synthesis		3D reconstruction	
	PSNR	SSIM	PSNR	SSIM
TensoRF	34.07	0.9830	22.82	0.9730
NAF	40.05	0.9962	23.97	0.9856
IntraTomo	37.69	0.9939	23.76	0.9852
NeRF	43.14	0.9982	20.99	0.7284
AME-Net	45.19	**0.9985**	25.11	0.9880
AME-Net++	**50.11**	**0.9985**	**26.88**	**0.9918**

The dataset comprises data from 30 experimental animals, including intraoperative angiographic projection images acquired from 10 views and 3D vascular reference data manually extracted from CTA images. Figure 4 shows a randomly selected example from the dataset. The best results are shown in bold, and the second-best results are underlined. PSNR: Peak signal-to-noise ratio; SSIM: structural similarity index measure; NAF: neural attenuation fields; NeRF: neural radiance fields; CTA: computed tomography angiography.

Table 6. Estimated average reconstruction time per case under the evaluated protocol using an NVIDIA A100 GPU

Method	Estimated reconstruction time	Hardware
IntraTomo	40.0 min	NVIDIA A100 GPU
NeRF	6.56 h	NVIDIA A100 GPU
FDK	23.3 min	NVIDIA A100 GPU
TensoRF	1.28 h	NVIDIA A100 GPU
NAF	1.17 h	NVIDIA A100 GPU
AME-Net++	30.39 h	NVIDIA A100 GPU

For neural reconstruction methods, the reported time corresponds to the fixed 2,000-epoch optimization protocol on the public CT datasets. For FDK, the reported time corresponds to analytical reconstruction and does not involve optimization epochs. NeRF: Neural radiance fields; FDK: Feldkamp–Davis–Kress; NAF: neural attenuation fields; CT: computed tomography.

On the chest scene, AME-Net++ achieves 50.35/0.9995, outperforming NAF (42.37/0.9993). On the leg scene, it reaches 59.46/0.9998, achieving the highest reported values for both metrics. PSNR improvements over AME-Net are also observed on the foot, head, pancreas, abdomen, and aneurysm scenes. In particular, on the head scene, AME-Net++ achieves 55.69/0.9997, whereas the external baseline with the highest PSNR reaches 34.83/0.9977. On the aneurysm scene, AME-Net++ achieves the highest PSNR of 54.76 dB, although its SSIM of 0.9996 is slightly lower than those of AME-Net (0.9998) and TensoRF (0.9997). Overall, AME-Net++ achieves the highest average PSNR and SSIM among the compared methods, supporting its effectiveness for novel-view projection synthesis under the evaluated sparse-view setting.

Effect of the proposed design. The quantitative results show the combined contribution of the proposed architectural and loss designs. AME-Net consistently outperforms existing neural reconstruction baselines across most scenes. This improvement suggests that the gated mixture-of-experts design effectively models region-dependent attenuation patterns, especially in anatomically heterogeneous regions where a single shared representation may struggle to capture complex local variations.

Second, the proposed multi-scale information-aware weighted composite loss further strengthens the base model. In 3D reconstruction, it improves the average PSNR from 35.06 to 35.71 and the average SSIM from 0.9688 to 0.9732. In projection synthesis, it further boosts the average PSNR from 50.88 to 52.88. These results indicate that the proposed loss enhances supervision over informative structures and edge-rich regions, leading to better global structural consistency and finer local detail preservation.

The qualitative comparisons in Figures 2 and 3 further support the quantitative findings. Compared with competing methods, AME-Net++ reconstructs clearer anatomical structures, introduces fewer artifacts, and preserves tissue boundaries more continuously. These advantages become especially prominent in regions with complex geometry, fine-scale structures, and heterogeneous attenuation patterns, where existing methods often suffer from blurred boundaries, missing details, or inconsistent intensities. Together, these results demonstrate the practical value of AME-Net++ for sparse-view projection-based volumetric reconstruction.

Ablation study. In the ablation study presented in Table 4, we conducted a progressive analysis of the key components of the proposed method. Starting from the baseline model, we sequentially integrated the MoE module and the sequence module to form AME-Net. Then we incorporated the entropy loss, local gradient-aware loss, and multi-scale loss to obtain the final AME-Net++. The ablation results show that the complete AME-Net++ configuration outperforms the baseline in both PSNR and SSIM, achieving the highest reported values among the tested configurations. However, the changes are not consistent across metrics, and SSIM

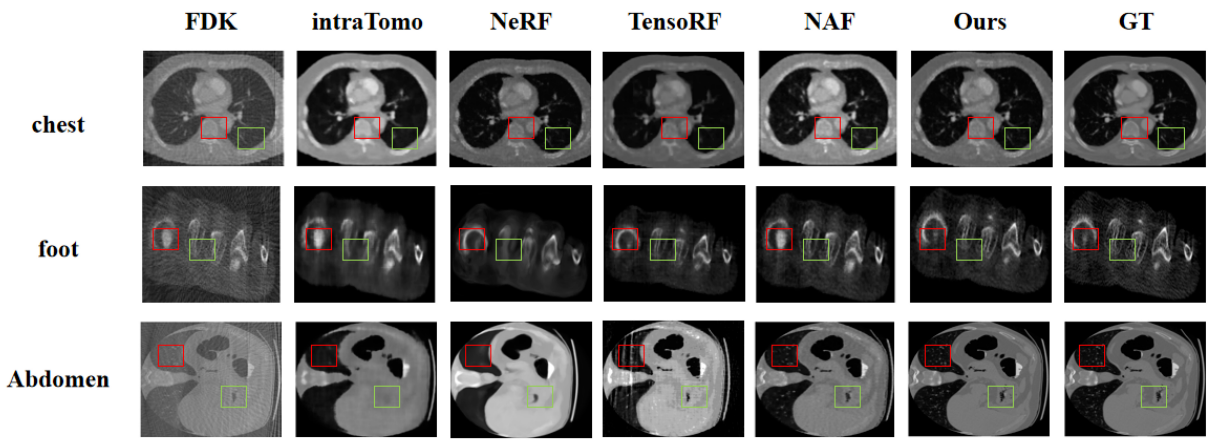


Figure 2. Comparison of reconstruction results on public datasets. The visualization demonstrates that AME-Net++ better preserves anatomical boundaries and structural continuity under sparse-view reconstruction. FDK: Feldkamp–Davis–Kress; NeRF: neural radiance fields; NAF: neural attenuation fields; GT: ground truth.

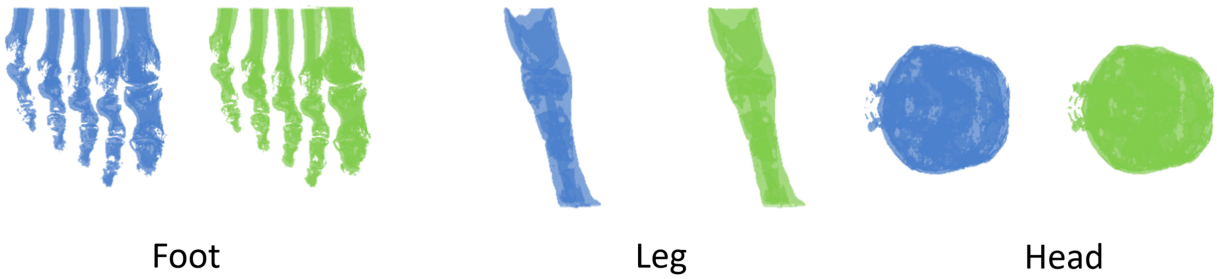


Figure 3. Comparison of multi-angle projection synthesis results with the corresponding ground truth on public datasets. Projections synthesized by our method are shown in blue, and ground-truth projections are shown in green for three anatomical regions: foot, leg, and head.

does not increase with every component addition. Specifically, SSIM remains at 0.9993 at the reported precision after adding the entropy loss and the local gradient-aware loss, and increases to 0.9994 after adding the multi-scale loss. Overall, these results support the effectiveness of the combined design in improving overall model performance.

Validation on the self-built animal interventional vascular dataset. Table 5 reports the quantitative reconstruction results on the self-built animal interventional vascular dataset. The proposed AME-Net++ achieves stable vascular reconstruction performance and preserves the overall vascular morphology in most cases. As shown in Figure 4, the qualitative reconstruction results on this dataset further demonstrate good structural continuity and spatial consistency with the reference vasculature. These results in an animal interventional setting suggest that AME-Net++ is not limited to public or simulated datasets and support its potential applicability to vascular imaging data acquired during animal interventional experiments.

Beyond voxel-wise metrics (PSNR and SSIM), we performed a centerline-based geometric evaluation of the reconstructed vasculature. The centerline error was measured using the mean centerline distance, defined as the average closest-point Euclidean distance between the reconstructed vascular centerline and the CTA-derived reference centerline. Under this metric, AME-Net++ achieved a centerline error of less than 3 mm in the evaluated 10-view setting, indicating its ability to preserve the primary vascular trajectories with good spatial accuracy. The qualitative results also show close agreement with the ground truth in overall vascular

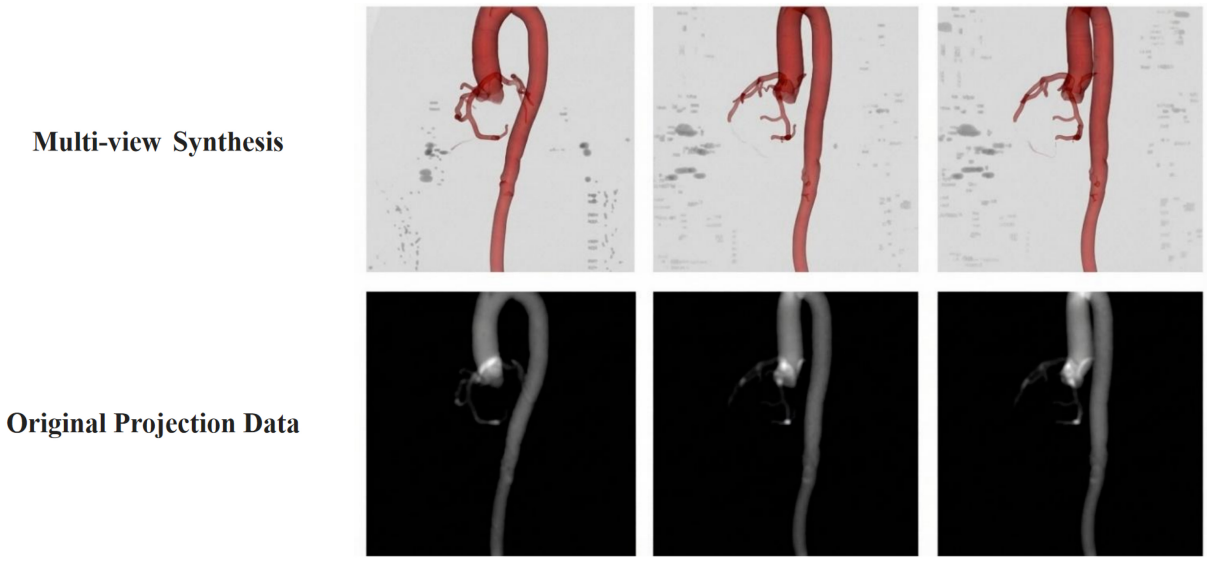


Figure 4. Visualization of original vascular projection data and multi-view synthesis results from different viewing angles on our self-built animal interventional vascular dataset. The original projection data serve as the reference input, while the synthesized projections show consistent vascular morphology across multiple viewpoints.

course, branching structure, and spatial configuration, with no obvious artifacts or disruptions in the visualized cases.

Reconstruction time and optimization analysis. To evaluate the computational cost of AME-Net++, we compared the estimated average reconstruction time per case under the same hardware environment using an NVIDIA A100 GPU. As summarized in Table 6, the neural reconstruction methods were evaluated under a fixed optimization protocol of 2,000 epochs, whereas FDK was evaluated as an analytical reconstruction method without neural-network optimization.

Under this fixed protocol, AME-Net++ required an estimated average reconstruction time of 30.39 h on the public CT datasets, which was longer than those of the compared methods. The additional computational cost is associated with sequence-based feature processing, gated mixture-of-experts modeling, and multi-scale loss evaluation. Although Top-1 routing evaluates only one expert for each feature token, the contextual processing, routing, and expert-dispatch operations still introduce computational overhead. Thus, AME-Net++ should currently be regarded as a reconstruction-quality-oriented framework rather than a real-time reconstruction method.

To complement the fixed-protocol runtime comparison, Figure 5 presents the evolution of 3D reconstruction PSNR and SSIM at checkpoints spaced every 100 epochs. The upper row shows the mean results across the seven public CT scenes, while the lower row shows the mean results across the 30 cases in the self-built animal interventional vascular dataset. PSNR and SSIM are presented in the left and right columns, respectively. At each checkpoint, the reconstructed volumes were evaluated using the same preprocessing and volumetric metric implementation described in Section 4.2. These curves characterize reconstruction quality during scene-specific optimization rather than cross-subject prediction performance.

The runtime required to complete the fixed protocol should be distinguished from the time required to reach a predefined reconstruction-quality criterion. In a separate retrospective assessment, approximately 1 h for public CT reconstruction and 4 min for animal vascular reconstruction corresponded to the elapsed time at

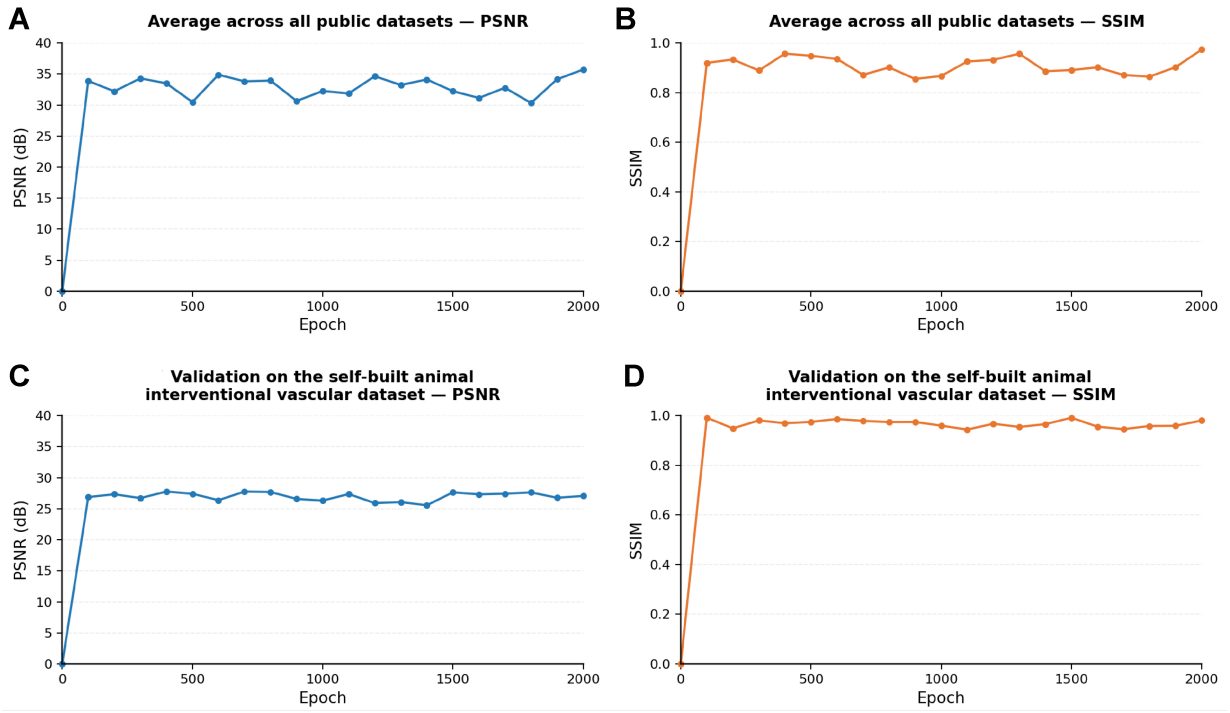


Figure 5. Evolution of 3D reconstruction quality during scene-specific optimization of AME-Net++. (A) Mean PSNR and (B) mean SSIM across the seven public CT scenes using 50 projection views per scene; (C) Mean PSNR and (D) mean SSIM across the 30 cases in the self-built animal interventional vascular dataset using 10 intraoperative angiographic views per case. Metrics were evaluated on reconstructed 3D volumes at checkpoints spaced every 100 epochs. The same axis ranges are used for corresponding metrics across the two datasets. PSNR: Peak signal-to-noise ratio; SSIM: structural similarity index measure; CT: computed tomography.

the first evaluated checkpoint where the relative discrepancy from the corresponding 2,000-epoch reference reconstruction was within 5%. These reference runs were used only for retrospective assessment and did not determine optimization stopping.

The public CT and animal vascular experiments differ in both reconstruction targets and acquisition settings: the former reconstructs heterogeneous full volumes from 50 projection views per scene, whereas the latter focuses on vascular structures from 10 intraoperative angiographic views per case. Their time-to-quality values are therefore setting-specific and should not be interpreted as a controlled comparison of computational efficiency. Moreover, agreement with a reference reconstruction does not establish clinical acceptability, and the retrospective criterion is not an intraoperative stopping rule. Together, the runtime and checkpoint analyses provide complementary descriptions of computational cost and reconstruction quality, while highlighting the need for further acceleration and prospective validation.

Clinical utility of reconstructed volumes. Regarding downstream utility, AME-Net++ is not mainly designed as a preprocessing module for conventional segmentation or classification tasks. Instead, its primary objective is to reconstruct intraoperative 3D anatomical information from a limited number of X-ray projections, thereby providing a 3D visual representation for image-guided intervention, robotic surgical navigation, and intraoperative spatial mapping. In current clinical practice, physicians usually observe vascular structures through only a few 2D X-ray projections and need to infer the underlying 3D spatial anatomy from these limited views based on substantial anatomical experience and spatial imagination, which imposes a high cognitive burden. Although contrast-enhanced angiographic projections provide important intraoperative information, obtaining a true intraoperative 3D vascular reference remains difficult. The reconstructed 3D volume can help physicians observe vascular morphology from multiple virtual viewpoints, better understand the global vascular contour and spatial course, and support navigation-oriented

intraoperative decision-making. Therefore, the clinical value of the proposed method lies primarily in providing an intraoperative 3D visual backbone rather than improving a single conventional downstream recognition task.

6. DISCUSSION

Although AME-Net++ achieves improved reconstruction performance under the evaluated sparse-view settings, several limitations should be acknowledged.

First, the required number of projection views cannot yet be reduced to the extremely low level ideally expected in practical intraoperative scenarios. Clinically, it is desirable to complete reconstruction from only a few projection angles to reduce radiation exposure, shorten acquisition time, and improve intraoperative workflow. However, under such ultra-sparse-view conditions, the available geometric constraints are severely insufficient, which may lead to structural ambiguity, missing fine details, and degraded reconstruction fidelity.

In this study, the 50-view setting is mainly used as a controlled sparse-view benchmark for complex full-volume reconstruction involving multiple anatomical structures. For clinically oriented vascular reconstruction on our self-built animal interventional vascular dataset, we used a more realistic 10-view limited-angle setting. The results suggest that relatively sparse and concentrated vascular structures can be reconstructed from fewer views than complex multi-tissue CT volumes. Nevertheless, more extreme settings, such as 5-view reconstruction, remain challenging and require further investigation.

Second, the advantage of AME-Net++ depends on the complexity of the reconstruction target. The proposed method is mainly designed for complex reconstruction scenarios involving heterogeneous attenuation distributions, weak contrast regions, and region-dependent anatomical variations. When the target structure is relatively simple, such as isolated vascular reconstruction with clear geometry and limited surrounding tissue interference, simpler NeRF-based methods may already achieve competitive results. In such cases, the advantage of the adaptive mixture-of-experts design may become less pronounced. The strength of AME-Net++ is more evident in anatomically complex scenarios where multiple tissue types, weak contrast, and heterogeneous attenuation patterns coexist.

Third, AME-Net++ is not a conventional supervised model with a fixed training/testing split. Instead, it follows a NeRF-style implicit representation paradigm, whose goal is to establish the mapping relationship between projection observations and the corresponding 3D spatial representation for each scene or patient. Since each patient has different anatomical space and vascular morphology, a new mapping relationship needs to be established for each patient based on the acquired projection images. Therefore, the current evaluation focuses on reconstruction quality under sparse-view conditions rather than cross-patient generalization in the conventional supervised-learning sense.

Fourth, computational efficiency remains an important limitation. For complex high-resolution CT reconstruction, AME-Net++ requires longer reconstruction time than several baseline methods because of the sequence modeling module, gated mixture-of-experts routing, and multi-scale information-aware optimization. Therefore, AME-Net++ should currently be positioned as a reconstruction-quality-oriented framework rather than a real-time reconstruction method. For vascular-only reconstruction, the optimization converges much faster due to the sparse and high-contrast nature of the target structure, but further acceleration is still necessary for routine intraoperative deployment.

The reported reconstruction times represent two distinct measures: the runtime of a fixed optimization protocol and a retrospective time-to-quality measure. The 30.39 h reported for AME-Net++ in Table 6 corresponds to completion of the fixed 2,000-epoch protocol used for neural-method comparisons on the public CT datasets. In contrast, approximately 1 h for full-volume reconstruction on the public CT datasets covering seven anatomical regions and 4 min for vascular-only reconstruction on our in-house animal interventional vascular dataset denote the elapsed time at the first evaluated checkpoint where the relative discrepancy from the corresponding 2,000-epoch reference reconstruction was within 5%. These reference runs were used solely for retrospective assessment and did not determine when optimization was stopped. The public CT benchmark used 50 input projection views per scene, whereas the in-house animal vascular experiments used 10 intraoperative angiographic views per case. Given these differences in input views and reconstruction targets, the time-to-quality values are setting-specific and not directly comparable. Moreover, agreement with a reference reconstruction does not establish clinical acceptability, and this retrospective criterion is not an intraoperative stopping rule. These findings suggest potential for faster vascular reconstruction but do not establish real-time applicability.

Fifth, the loss weights and other major hyperparameters were manually predefined before the full experiments and were then kept unchanged across all datasets, anatomical scenes, and experimental cases. No particular scene or subject was used for scene-specific parameter selection, and no systematic grid search or other automated hyperparameter optimization was performed. Therefore, the reported configuration should be interpreted as a fixed practical setting rather than a globally optimal parameter combination. Future work may investigate Bayesian optimization, adaptive loss-weight learning, or uncertainty-based weighting to reduce manual parameter specification.

Finally, although the current framework improves reconstruction through adaptive feature modeling, gated mixture-of-experts representation, and multi-scale information-aware supervision, it does not yet explicitly incorporate strong prior knowledge, such as anatomical shape priors, vessel topology constraints, pretrained neural field priors, or generative reconstruction priors. Without these priors, it remains challenging to substantially reduce the number of required projection views while maintaining high reconstruction accuracy.

Future work will focus on more clinically practical ultra-sparse-view reconstruction settings and on incorporating stronger prior information into the reconstruction framework. Potential directions include anatomical priors, vascular topology constraints, pretrained shape or neural representation priors, diffusion-based generative priors, temporal consistency from sequential intraoperative X-ray images, and more efficient patient-specific mapping construction. We believe that combining adaptive neural reconstruction with such priors is necessary for further reducing the required number of projections and improving the clinical applicability of intraoperative 3D reconstruction.

7. CONCLUSION

This paper investigates intraoperative 3D reconstruction from multi-angle X-ray projections under sparse-view conditions. To tackle the severe information degradation, compromised local detail fidelity, and reconstruction instability, we propose an adaptive mixture-of-experts network (AME-Net++). Our approach incorporates a local-global collaborative modeling strategy that couples sequence-dependency learning along ray trajectories with a gated Mixture-of-Experts mechanism, thereby achieving region-specific attenuation representation and reconstruction. Furthermore, a multi-scale awareness composite loss is designed to concurrently enhance global anatomical consistency and local detail fidelity. Extensive experiments demonstrate that AME-Net++ achieves superior performance in both projection synthesis and 3D volumetric reconstruction across various medical imaging benchmarks, while maintaining high robustness

under sparse-view constraints. Statistical analyses provide exploratory support for improvements in selected metrics, while other gains remain numerical rather than statistically significant. Crucially, validation on an in-house animal interventional vascular dataset confirms that the framework reliably preserves global vessel morphology and principal branches, keeping the mean centerline error within 3 mm. As a framework tailored for intraoperative interventions, this paradigm demonstrates strong clinical potential for facilitating low-radiation 3D navigation.

DECLARATIONS

Acknowledgments

The authors would like to express their sincere gratitude to Prof. Shao-Feng Guan, and Dr. Ming Wang from Huadong Hospital for their invaluable clinical guidance and insightful discussions. We also thank all members of the research team for their continuous technical assistance and helpful discussions throughout this work.

Authors' contributions

Conceptualization, methodology, algorithm design, experiments, data analysis, and manuscript writing: Liu, S. Q.
Animal interventional data support, experimental validation, and manuscript revision: Zhang, Y. J.
Methodology discussion, software implementation, and result analysis: Xie, X. L.
Data preprocessing and experimental support: Zhang, L. S.
Technical support and manuscript revision: Zhou, X. H.
Project discussion and manuscript revision: Ma, X. Y.
Supervision, project administration, and funding acquisition: Hou, Z. G.
Clinical and experimental supervision, data curation, project administration, and manuscript revision: Qu, X. K.

Availability of data and materials

The public datasets used in this study can be accessed from the following sources: LIDC-IDRI[35] is available through The Cancer Imaging Archive (TCIA) at https://www.cancerimagingarchive.net/collection/lidc-idri/; the MIDA anatomical model[36] can be obtained through the IT'IS Foundation at https://itis.swiss/virtual-population/regional-human-models/mida-model/, subject to the applicable license agreement and access procedure; and the Pancreas-CT dataset associated with DeepOrgan[37] is available through TCIA at https://www.cancerimagingarchive.net/collection/pancreas-ct/, with download instructions provided on the collection page. Source links and download instructions for the VOLVIS datasets[38] are provided in the Teem documentation at https://teem.sourceforge.net/nrrd/volvis/index.html. Users should follow the applicable access requirements, licenses, and citation policies specified by the respective data providers.

AI and AI-assisted tools statement

Not applicable.

Financial support and sponsorship

This work was supported in part by the National Key Research and Development Program of China (Grant 2024YFF1206902); in part by the Development Project of National Major Scientific Research Instrument (Grant 82327801); in part by the National Natural Science Foundation of China (Grant 62303463); in part by the Shanghai Municipal Key Clinical Specialty (Grant shslczdzk02801); and in part by the Beijing Natural Science Foundation (Grants L246047, L232137, and Z241100009024031).

Conflicts of interest

Hou, Z. G. serves as a Section Editor for the Robot Control and Navigation section of the journal *Intelligence & Robotics* but was not involved in any stage of the editorial processing of this manuscript, including reviewer selection, manuscript handling, or decision-making. The other authors declare that they have no conflicts of interest.

Ethical approval and consent to participate

All animal experiments, including the imaging and experimental procedures used to establish the in-house animal interventional vascular dataset, were conducted at the Laboratory Animal Center of Shanghai Xinova Medical Research Co., Ltd. and were approved by the company's Animal Ethics Committee (approval No. XNM-YX-20221103-01). The center holds a laboratory animal production license [No. SCXK (Su) 2021-0007] and a laboratory animal use license [No. SYXK (Hu) 2021-0019]. All animal experiments were conducted in accordance with institutional and national guidelines for the care and use of laboratory animals.

Consent for publication

Not applicable.

Copyright

© The Author(s) 2026.

REFERENCES

1. Zhang, B.; Liu, S.; Xie, X.; et al. HDCAR: a 3D-2D registration network for abdominal aortic vessels based on CTA vessel models and DSA images. *Biomim Intell Robot* **2026**, *6*, 100272. DOI
2. Feldkamp, L. A.; Davis, L. C.; Kress, J. W. Practical cone-beam algorithm. *J Opt Soc Am A* **1984**, *1*, 612-9. DOI
3. Andersen, A. H.; Kak, A. C. Simultaneous algebraic reconstruction technique (SART): a superior implementation of the ART algorithm. *Ultrason Imaging* **1984**, *6*, 81-94. DOI
4. Shin, H.; Kim, T.; Lee, J.; Chun, S. Y.; Cho, S.; Shin, D. Sparse-view CBCT reconstruction using meta-learned neural attenuation field and hash-encoding regularization. *Comput Biol Med* **2025**, *189*, 109900. DOI PubMed
5. Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; Ng, R. NeRF: representing scenes as neural radiance fields for view synthesis. *Commun ACM* **2022**, *65*, 99-106. DOI
6. Gordon, R.; Bender, R.; Herman, G. T. Algebraic reconstruction techniques (ART) for three-dimensional electron microscopy and X-ray photography. *J Theor Biol* **1970**, *29*, 471-81. DOI PubMed
7. Huang, J.; Zhang, Y.; Ma, J.; et al. Iterative image reconstruction for sparse-view CT using normal-dose image induced total variation prior. *PLoS One* **2013**, *8*, e79709. DOI PubMed PMC
8. Niu, S.; Gao, Y.; Bian, Z.; et al. Sparse-view x-ray CT reconstruction via total generalized variation regularization. *Phys Med Biol* **2014**, *59*, 2997-3017. DOI PubMed PMC
9. Huang, Y.; Taubmann, O.; Huang, X.; Haase, V.; Lauritsch, G.; Maier, A. Scale-space anisotropic total variation for limited angle tomography. *IEEE Trans Radiat Plasma Med Sci* **2018**, *2*, 307-14. DOI
10. Xu, M.; Hu, D.; Luo, F.; Liu, F.; Wang, S.; Wu, W. Limited-angle X-ray CT reconstruction using image gradient ℓ_0-norm with dictionary learning. *IEEE Trans Radiat Plasma Med Sci* **2021**, *5*, 78-87. DOI
11. Rückert, D.; Wang, Y.; Li, R.; Idoughi, R.; Heidrich, W. NeAT: neural adaptive tomography. *ACM Trans Graph* **2022**, *41*, 1-13. DOI
12. Chen, A.; Xu, Z.; Geiger, A.; Yu, J.; Su, H. TensoRF: tensorial radiance fields. In *Computer Vision - ECCV 2022*. Cham: Springer Nature Switzerland; Avidan, S.; Brostow, G.; Cissé, M.; Farinella, G. M.; Hassner, T.; Eds.; Springer, Cham.; 2022. pp. 333-50. DOI
13. Zha, R.; Zhang, Y.; Li, H. NAF: neural attenuation fields for sparse-view CBCT reconstruction. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2022*. Wang, L.; Dou, Q.; Fletcher, P. T.; Speidel, S.; Li, S; Eds.; Springer, Cham; 2022. pp. 442-52. DOI
14. Zang, G.; Idoughi, R.; Li, R.; Wonka, P.; Heidrich, W. IntraTomo: self-supervised learning-based tomography via sinogram synthesis and prediction. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, Canada. Oct 10-17, 2021. IEEE; 2021. pp. 1960-70. DOI
15. Sidky, E. Y.; Pan, X. Image reconstruction in circular cone-beam computed tomography by constrained, total-variation minimization. *Phys Med Biol* **2008**, *53*, 4777. DOI
16. Barron, J. T.; Mildenhall, B.; Tancik, M.; Hedman, P.; Martin-Brualla, R.; Srinivasan, P. P. Mip-NeRF: a multiscale representation for anti-aliasing neural radiance fields. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, Canada. Oct 10-17, 2021. IEEE; 2021. pp. 5855-64. DOI
17. Wang, S.; Cheng, Q.; Gasperini, S.; et al. VoxNeRF: bridging voxel representation and neural radiance fields for enhanced indoor view synthesis. *IEEE Robot Autom Lett* **2025**, *10*, 5903-10. DOI
18. Irshad, M. Z.; Zakharov, S.; Guizilini, V.; Gaidon, A.; Kira, Z.; Ambrus, R. NeRF-MAE: masked AutoEncoders for self-supervised 3D representation learning for neural radiance fields. In *Computer Vision - ECCV 2024*, Milan, Italy. Sep 29 - Oct 4, 2024. Cham: Springer Nature Switzerland; 2024. pp. 434-53. DOI

19. Wang, X.; Chen, Y.; Hu, S.; Fan, H.; Zhu, H.; Li, X. Neural radiance fields in medical imaging: a survey. *arXiv* **2024**, arXiv:2402.17797. Available online: https://doi.org/10.48550/arXiv.2402.17797. (accessed 2026-10-29).
20. Cai, Y.; Wang, J.; Yuille, A.; Zhou, Z.; Wang, A. Structure-aware sparse-view X-ray 3D reconstruction. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, USA. Jun 16-22, 2024. IEEE; 2024. pp. 11174-83. DOI
21. Lin, Y.; Wang, H.; Chen, J.; Li, X. Learning 3D Gaussians for extremely sparse-view cone-beam CT reconstruction. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2024*. Springer Nature Switzerland; 2024. pp. 425-35. DOI
22. Li, Y.; Fu, X.; Li, H.; Zhao, S.; Jin, R.; Zhou, S. K. 3DGR-CT: sparse-view CT reconstruction with a 3D Gaussian representation. *Med Image Anal* **2025**, *103*, 103585. DOI
23. Xia, C.; Gu, T.; Zheng, N.; Wei, H.; Tsai, T. Y. RNAF: regularization neural attenuation fields for sparse-view CBCT reconstruction. *J Xray Sci Technol* **2025**, *33*, 713-25. DOI PubMed
24. Müller, T.; Evans, A.; Schied, C.; Keller, A. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans Graph* **2022**, *41*, 1-15. DOI
25. Kerbl, B.; Kopanas, G.; Leimkuehler, T.; Drettakis, G. 3D Gaussian splatting for real-time radiance field rendering. *ACM Trans Graph* **2023**, *42*, 1-14. DOI
26. Fridovich-Keil, S.; Yu, A.; Tancik, M.; Chen, Q.; Recht, B.; Kanazawa, A. Plenoxels: radiance fields without neural networks. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, USA. Jun 18-24, 2022. IEEE; 2022. pp. 5491-500. DOI
27. Sun, C.; Sun, M.; Chen, H. T. Direct voxel grid optimization: super-fast convergence for radiance fields reconstruction. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, USA. Jun 18-24, 2022. IEEE; 2022. pp. 5449-59. DOI
28. Barron, J. T.; Mildenhall, B.; Verbin, D.; Srinivasan, P. P.; Hedman, P. Mip-NeRF 360: unbounded anti-aliased neural radiance fields. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, USA. Jun 18-24, 2022. IEEE; 2022. pp. 5460-9. DOI
29. Jacobs, R. A.; Jordan, M. I.; Nowlan, S. J.; Hinton, G. E. Adaptive mixtures of local experts. *Neural Comput* **1991**, *3*, 79-87. DOI
30. Wang, X.; Yu, F.; Dunlap, L.; et al. Deep mixture of experts via shallow embedding. In *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*. PMLR; 2020. pp. 552-62. https://proceedings.mlr.press/v115/wang20d.html. (accessed 2026-10-29).
31. Bengio, E.; Bacon, P. L.; Pineau, J.; Precup, D. Conditional computation in neural networks for faster models. *arXiv* **2015**, arXiv:1511.06297. Available online: https://doi.org/10.48550/arXiv.1511.06297. (accessed 2026-10-29).
32. Lepikhin, D.; Lee, H.; Xu, Y.; et al. GShard: scaling giant models with conditional computation and automatic sharding. *arXiv* **2020**, arXiv:2006.16668. Available online: https://doi.org/10.48550/arXiv.2006.16668. (accessed 2026-10-29).
33. Fedus, W.; Zoph, B.; Shazeer, N. Switch transformers: scaling to trillion parameter models with simple and efficient sparsity. *J Mach Learn Res* **2022**, *23*, 5232-70. DOI
34. Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, USA. Jun 18-23, 2018. IEEE; 2018. pp. 586-95. DOI
35. Armato, S. G. 3rd.; McLennan, G.; Bidaut, L.; et al. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans. *Med Phys* **2011**, *38*, 915-31. DOI PubMed PMC
36. Iacono, M. I.; Neufeld, E.; Akinnagbe, E.; et al. MIDA: a multimodal imaging-based detailed anatomical model of the human head and neck. *PLoS One* **2015**, *10*, e0124126. DOI PubMed PMC
37. Roth, H. R.; Lu, L.; Farag, A.; et al. DeepOrgan: multi-level deep convolutional networks for automated pancreas segmentation. In *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015*. Navab, N.; Hornegger, J.; Wells, W. M.; Frangi, A.; Eds.; Cham: Springer International Publishing; 2015. pp. 556-64. DOI
38. Teem. Inspecting volvis.org datasets. https://teem.sourceforge.net/nrrd/volvis/index.html. (accessed 2026-10-29).

Disclaimer/Publisher's Note: All statements, opinions, and data contained in this publication are solely those of the individual author(s) and contributor(s) and do not necessarily reflect those of OAE and/or the editor(s). OAE and/or the editor(s) disclaim any responsibility for harm to persons or property resulting from the use of any ideas, methods, instructions, or products mentioned in the content.

© The Author(s) 2026. Open Access This article is licensed under a Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, sharing, adaptation, distribution and reproduction in any medium or format, for any purpose, even commercially, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.