3DABSeg: Adaptive 3D Ankle Bone Segmentation with Multiscale Feature Fusion Mixture-of-Experts

Xiaoling Luo[1] **Yuhan Tang**[2,4] **Yang Zhang**[1*] **Tianxing Zhao**[1] **Xiaowen Fu**[2]
Xiaowen Huang[1] **Hongbo Le**[3] **Xi Zhou**[3] **Linlin Shen**[2]

[1] College of Computer Science and Software Engineering, Shenzhen University
[2] School of Artificial Intelligence, Shenzhen University
[3] Department of Radiology, South China Hospital, Medical School, Shenzhen University
[4] Intelligent Computing Center, Shenzhen University

Abstract

As one of the most complex and load-bearing joints in the human body, the ankle plays a crucial role in locomotion and clinical assessment. Accurate segmentation of ankle bones is essential for trauma evaluation, preoperative planning, and disease diagnosis. However, the scarcity of publicly available ankle datasets has hindered the development of intelligent analysis in this area. To address this gap, we construct two high-quality ankle CT datasets, providing new directions for medical image segmentation. Considering the intricate and spatially correlated anatomy of the ankle, we propose a 3D medical image segmentation model, namely 3DABSeg. Specifically, we introduce the Hybrid KAN-Mamba Block (HKMB) to capture long-range spatial dependencies and complex nonlinearities. By integrating Mamba's sequence modeling with KAN's learnable spline functions, HKMB enhances feature expressiveness for complex anatomical structures. Furthermore, to address the uneven semantic distribution and feature dilution across channels, we propose a Multi-scale Feature Fusion Mixture-of-Experts module (MFF-MoE). MFF-MoE utilizes multi-scale spatial pooling to compute dynamic routing scores, adaptively partitioning channels into specialized expert networks via a mutually exclusive hard-assignment strategy. Extensive experiments demonstrate that 3DABSeg surpasses existing state-of-the-art methods in both segmentation performance and boundary precision, highlighting its potential for complex anatomical structure analysis and clinical applications. The project source code is publicly available at `https://github.com/Yikol/3DABSeg`.

1 Introduction

Medical image segmentation plays a fundamental role in modern healthcare by enabling precise delineation of anatomical structures and pathological regions to support accurate diagnosis, treatment planning, and surgical navigation [38]. As one of the most complex and load-bearing joints, the ankle comprises multiple small bones with intricate morphologies and dense spatial relationships. Accurate segmentation of ankle bones is essential for trauma assessment, fracture diagnosis, and postoperative rehabilitation. However, the lack of publicly available ankle datasets has significantly limited research in this area. To address this gap, we construct two high-quality ankle CT segmentation datasets, providing a new direction for future studies on medical image segmentation.

Compared with other anatomical regions, the ankle exhibits substantial variations in bone scale, fuzzy and irregular boundaries, significant intensity fluctuations, and strong coupling between local details

*Corresponding author: Yang Zhang

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

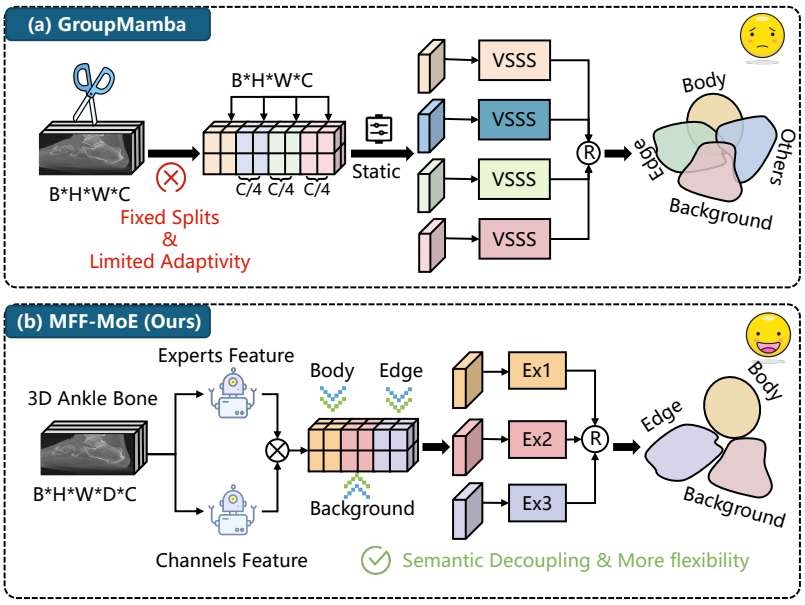

Figure 1: (a) The modulated Group Mamba layer of GroupMamba [34]. (b) The proposed MFF-MoE module in 3DABSeg.

and global structures. Traditional convolutional neural networks (CNNs) such as U-Net [31] and UNet++ [48] excel at capturing local anatomical details and boundary features, achieving remarkable performance on liver segmentation tasks [19]. However, their limited receptive fields make it difficult to model long-range dependencies between bones. Transformer-based architectures [5, 12] enhance global modeling through self-attention [37] and have shown strong performance in multimodal segmentation tasks. However, their quadratic computational complexity and reliance on large scale data limit their applicability in practical 3D medical scenarios. In contrast, Mamba [10, 9] maintains the ability to model global dependencies with linear complexity, making it suitable for ankle CT segmentation characterized by intricate geometry and tightly coupled spatial structures. However, its scanning mechanism across patches struggles to capture the non-linear dependencies and complex semantic interactions among channels in medical images.

To address these challenges, we propose 3DABSeg, a 3D ankle bone segmentation network that integrates efficient spatial modeling with adaptive channel grouping. While Mamba-based models have demonstrated exceptional efficiency in capturing long-range spatial dependencies, their scanning mechanisms focus on spatial serialization, leaving inter-channel interactions dependent on fixed linear projections. However, intensity variations at bone boundaries in ankle CT images are highly complex, and such static linear projections lack the flexibility to dynamically adjust mapping relationships according to input features. Inspired by the "global observation, local focus[3]" diagnostic strategy employed by radiologists, we design the Hybrid KAN-Mamba Block (HKMB). Specifically, Mamba's linear scanning is utilized to establish global spatial relationships, ensuring the accurate positioning of anatomical structures. Building upon this, the Kolmogorov-Arnold Network (KAN) is introduced as an adaptive non-linear mapper for the channel dimension, employing learnable spline functions to model complex inter-channel relationships. This design enables HKMB to adaptively amplify feature expressiveness in boundary regions, thereby enhancing the precision of edge segmentation.

Although HKMB enhances global spatial consistency and nonlinear channel modeling, it remains insufficient to cope with the inherently uneven semantic distribution across channels. As the network deepens, convolution and downsampling progressively scatter channel semantics, causing features of fine-grained bone structures to be diluted by dominant channels. Prior research [4] further shows that many channels in CNNs are redundant or highly correlated, indicating that raw channel features are neither equally informative nor semantically balanced. Traditional skip connections [6] directly concatenate encoder features without considering semantic discrepancies among channels, resulting in information entanglement. As illustrated in Figure 1 (a), GroupMamba [34] directly divides channels into four fixed groups along the channel dimension, overlooking potential semantic correlations. To

overcome this limitation, we propose the Multi-scale Feature Fusion Mixture-of-Experts (MFF-MoE) module, which adaptively performs fine-grained feature aggregation at the channel level, as shown in Figure 1 (b). This module extracts expert representations and channel features through multiscale spatial pooling operations to compute their respective dynamic assignment scores. By implementing a mutually exclusive hard assignment strategy, MFF-MoE adaptively routes channels to specialized expert networks, facilitating genuine semantic decoupling and allowing each expert to focus on a distinct semantic subspace. This content-dependent interaction mechanism enables refined feature modeling for complex anatomical structures, significantly enhancing the semantic discriminability and structural representation of channel features.

In summary, 3DABSeg integrates HKMB and MFF-MoE to capture global structures and fine-grained anatomical details. It simulates the radiologist's "global observation, local focus" strategy, achieving accurate segmentation with precise boundaries and enhanced robustness in ankle bone regions.

Our main contributions can be summarized as follows:

- We propose the HKMB to strengthen long range spatial reasoning and nonlinear channel interaction, enabling the model to progressively capture both global anatomical structures and fine-grained bone details. This design enhances feature representation quality and contributes to improving segmentation accuracy and boundary precision.
- We introduce the MFF-MoE module to adaptively group channel features into semantically coherent subspaces and route them to specialized experts, enabling fine-grained feature refinement and improving segmentation consistency in complex regions.
- We construct two 3D ankle bone segmentation datasets, providing high-quality annotations and inspiring new directions in 3D medical image segmentation.

2 Related Work

2.1 Medical Image Segmentation

Since the inception of U-Net [31] and its 3D extension [6], variants such as UNet++ [48], nnU-Net [14], and e2eNet [40] have refined medical image segmentation through encoder-decoder designs and enhanced feature fusion. To overcome the limited receptive field of CNNs, Transformer-based frameworks [12, 5, 39] were introduced to capture global context, albeit at the cost of high computational complexity. Recently, State Space Models (SSMs) [10] have emerged as efficient alternatives, with models such as U-Mamba [28], SegMamba [43], and LightM-UNet [20] integrating Mamba modules to balance local extraction with long-range modeling. Beyond segmentation, recent studies have employed multi-view Hilbert curve-based Mamba architectures to aggregate local and global information for incomplete multimodal Alzheimer's disease diagnosis [23, 22], further demonstrating the potential of SSMs for modeling high-dimensional heterogeneous medical data. However, despite their efficiency, the inherently linear formulation of Mamba models often limits their capacity to capture the complex nonlinear patterns and fine-grained anatomical details essential for precise volumetric segmentation.

2.2 Kolmogorov-Arnold Networks

Kolmogorov-Arnold Networks (KANs) [25], based on the Kolmogorov–Arnold representation theorem [17], replace fixed activations with learnable univariate mappings, offering superior interpretability and nonlinear flexibility. Recently, KANs have been integrated into various vision architectures to enhance feature representation. For instance, U-KAN [18], SegKAN [35], and KANDU-Net [7] embed KAN layers into U-Net to strengthen nonlinear expressivity through functional mappings or dual-channel designs. Other works, such as KM-UNet [46] and KAN-Mamba FusionNet [1], combine KAN's functional approximation with Mamba's efficient modeling for better local-global coordination. Furthermore, architectures like TransUKAN [41] and KAN-Mixer [8] leverage KAN for adaptive multi-scale modeling within Transformer or MLP-based frameworks. Collectively, these studies underscore KAN's adaptability and performance advantages across diverse visual tasks.

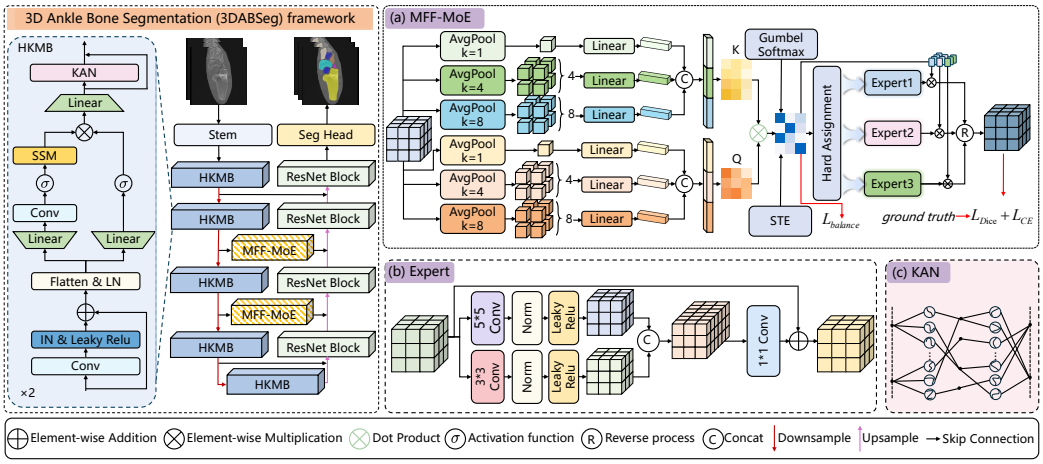

Figure 2: Overview of the proposed 3DABSeg framework with HKMB and MFF-MoE. (a) Detailed architecture of MFF-MoE. (b) Detailed architecture of an individual expert module in MFF-MoE. (c) Detailed architecture of the KAN.

3 Methodology

3.1 The Architecture

As illustrated in Figure 2, the proposed 3DABSeg model is based on a symmetrical U-shaped encoder-decoder architecture, comprising three primary components: a hierarchical encoder integrated with Hybrid KAN-Mamba Blocks, an adaptive skip connection mechanism powered by Multi-Scale Feature Fusion Mixture-of-Experts, and a symmetrical decoder based on residual blocks.

In the initial stage, the input 3D image is projected into an initial feature representation via a Stem module. The encoder consists of five stages designed to progressively extract multi-scale hierarchical features and compress spatial dimensions. At each encoding stage, we apply HKMB to capture long-range spatial dependencies and leverage the flexibility of KAN to model complex nonlinear interactions along the channel dimension. Subsequently, the features output by the encoder are either sent directly to the decoder or passed through the MFF-MoE module for adaptive enhancement. By semantically decoupling channel features, this module routes different channel groups to specialized experts for targeted processing, thereby strengthening the feature representation of complex anatomical structures. During the decoding stage, the model progressively restores spatial resolution through upsampling operations. Each decoder layer concatenates and fuses deeper features with those from the skip connections to precisely reconstruct anatomical structural information. Finally, a segmentation head outputs voxel-wise mask predictions. During the training process, we jointly apply Dice loss, cross-entropy loss, and an expert load-balancing loss to optimize the model.

3.2 Hybrid KAN-Mamba Block

Anatomical structures in medical images often exhibit highly complex and nonlinear characteristics, such as the fractal-like branching of vessels, irregular tumor boundaries, and blurred interfaces between adjacent organs. These intricate patterns make it challenging to simultaneously capture long-range dependencies and subtle nonlinear variations.

According to the Kolmogorov-Arnold representation theorem, any multivariate continuous function can be decomposed into a finite composition of univariate functions [25]. KANs leverage learnable spline functions to implement adaptive nonlinear transformations that dynamically adjust to the characteristics of specific tasks and datasets. To better model complex nonlinear patterns, we propose the HKMB. Specifically, Mamba first captures long range spatial dependencies through efficient linear modeling, producing a coarse global representation. Then, KAN performs nonlinear refinement along the channel dimension, enhancing feature diversity and enabling the representation of complex anatomical details.

As illustrated in Figure 2, given the input feature map from the i-th encoder stage $X_i \in \mathbb{R}^{B\times C_i\times D_i\times H_i\times W_i}$, we first flatten its spatial dimensions to obtain a voxel-ordered sequence $S_i \in \mathbb{R}^{B\times N_i\times C_i}$, where $N_i = D_i \times H_i \times W_i$. To ensure numerical stability and feature consistency, Layer Normalization is applied along the channel dimension. The HKMB first utilizes a Mamba layer to perform efficient 3D contextual scanning with linear computational complexity. This design is specifically intended to capture global dependencies while maintaining the semantic integrity of the channel features. To further refine the Mamba output, we introduce a KAN-based residual branch. Unlike traditional fixed activation functions, KAN leverages learnable spline functions to perform high-order nonlinear transformations along the channel dimension. The forward process is formulated as:

$$Z_i = \mathrm{Mamba}(S_i) + \mathrm{KAN}(\mathrm{Mamba}(S_i)). \tag{1}$$

The residual operation ensures that KAN can adaptively enhance Mamba features without compromising the stability of the global representation.

Overall, HKMB integrates the strengths of Mamba and KAN. Mamba performs patch-token state scanning across the 3D spatial layout to efficiently model global context, while KAN enhances channel-wise nonlinearity to refine local semantic structures. Structurally decoupled yet functionally complementary, these two components jointly provide rich contextual information and fine-grained semantic features for subsequent encoding and skip connections.

3.3 Multi-Scale Feature Fusion Mixture-of-Experts Module

In 3D medical image segmentation, CNNs attempt to extract multi-scale and multi-semantic features across diverse channels. However, conventional convolutional layers typically rely on homogeneous filters across all channels, inherently ignoring the semantic heterogeneity between them. For instance, while certain channels are specialized in delineating intricate, fine-grained structures such as edges, others are more attuned to capturing global morphological attributes. This undifferentiated processing limits the model's representational capacity and induces significant inter-channel redundancy [4].

The Mixture-of-Experts (MoE) module facilitates meticulous modeling of various features by adaptively routing them to specialized expert networks. Recently, several researchers have introduced MoE into the field of medical image segmentation. For instance, Mamba-HoME [30] employs a Hierarchical Soft Mixture-of-Experts to route patch tokens across multi-layered experts, striving for a balance between fine-grained details and global context. Similarly, UTANet [27] integrates encoder features from multiple levels into the TA-MoSC module within the skip connections, employing a soft routing mechanism to facilitate feature interaction and reorganization. However, soft assignment strategies blur the semantic boundaries between experts, making it difficult to achieve genuine decoupling. Furthermore, the load-balancing loss function in traditional MoE forces each expert to process an equivalent number of channels, which significantly distorts the distribution of semantic features.

To address these challenges, we propose a Mixture-of-Experts module based on multi-scale feature fusion. By implementing a dynamic and mutually exclusive hard-assignment mechanism, our approach partitions channels based on semantic similarity and dispatches them to specialized expert networks. Furthermore, we introduce a load balancing loss based on entropy that prevents routing collapse while minimally interfering with the natural distribution of semantic features.

As illustrated in Figure 2 (a), given an input $\mathbf{X} \in \mathbb{R}^{B\times C\times D\times H\times W}$. To extract hierarchical features, we first process $\mathbf{X}$ using three global pooling layers with varying spatial scales:

$$\begin{cases} q_1 = \phi_{q1}(AvgPool_1(\mathbf{X}).view(B,-1)) \in \mathbb{R}^{B\times \lfloor P/3 \rfloor}, \\ q_2 = \phi_{q2}(AvgPool_4(\mathbf{X}).view(B,-1)) \in \mathbb{R}^{B\times \lfloor P/3 \rfloor}, \\ q_3 = \phi_{q3}(AvgPool_8(\mathbf{X}).view(B,-1)) \in \mathbb{R}^{B\times (P-2\lfloor P/3 \rfloor)}, \end{cases} \tag{2}$$

where ϕ_q denotes the linear projection that maps the pooled features into the query space, and $AvgPool_i$ represents the adaptive average pooling operation with a scale factor of i. $P = G \times F$, where G and F denote the number of experts and the feature dimension per expert, respectively. The dimensionality of the query space for q_3 is specifically designed to accommodate cases where P is not divisible by 3, thereby ensuring the robustness of the model. Finally, q_1, q_2, and q_3 are concatenated and normalized to generate the final query representation $Q \in \mathbb{R}^{B\times G\times F}$:

$$Q = Norm\left([q_1; q_2; q_3]\right). \tag{3}$$

To generate the key feature $\mathbf{K}$, the input $\mathbf{X}$ is first reshaped to $\hat{\mathbf{X}} \in \mathbb{R}^{(B\times C)\times 1\times D\times H\times W}$. Mirroring the formulation of $\mathbf{Q}$, the generation process for $\mathbf{K}$ is defined as follows:

$$\begin{cases} k_1 = \phi_{k1}(AvgPool_1(\hat{\mathbf{X}}).view(B\times C,-1)) \in \mathbb{R}^{(B\times C)\times \lfloor F/3 \rfloor}, \\ k_2 = \phi_{k2}(AvgPool_4(\hat{\mathbf{X}}).view(B\times C,-1)) \in \mathbb{R}^{(B\times C)\times \lfloor F/3 \rfloor}, \\ k_3 = \phi_{k3}(AvgPool_8(\hat{\mathbf{X}}).view(B\times C,-1)) \in \mathbb{R}^{(B\times C)\times (F-2\lfloor F/3 \rfloor)}. \end{cases} \tag{4}$$

Similarly, ϕ_k denotes the linear projection operation responsible for mapping channel features into the key feature space. The final key representation $K \in \mathbb{R}^{B\times C\times F}$ is also obtained by concatenating k_1, k_2, and k_3, followed by a normalization layer.

To achieve a hard assignment while maintaining end-to-end differentiability, we employ the Gumbel-Softmax strategy [15] to compute the routing matrix. First, the raw similarity scores $S \in \mathbb{R}^{B\times C\times G}$ are obtained via the inner product of the query and key features $KQ^\top$. During training, to enhance the selection diversity and encourage exploration among experts, Gumbel noise is injected into the raw score matrix S:

$$\tilde{S}_{b,c,g} = \sigma\left(\frac{S_{b,c,g} + g_{b,c,g}}{\tau}\right), \tag{5}$$

where $g_{b,c,g} = -\log(-\log(u))$, $u \sim \text{Uniform}(0,1)$, represents the Gumbel noise samples, τ is the temperature coefficient that controls the smoothness of the resulting distribution and σ denotes the Softmax function. This strategy encourages the model to explore diverse expert combinations during the early stages of training, thereby preventing the system from trapping in suboptimal local optima or suffering from expert collapse. To ensure that each channel is processed by only one expert, we apply an argmax operation over the routing scores to identify the optimal expert for each channel and convert them into a one-hot representation:

$$\tilde{A}_{b,c,g} = \text{one_hot}\left(\underset{j\in\{1,\dots,G\}}{\text{argmax}}\ \tilde{S}_{b,c,j}\right), \tag{6}$$

where $\tilde{A} \in \{0,1\}^{B\times C\times G}$ denotes the final binary routing matrix. Specifically, $\tilde{A}_{b,c,g} = 1$ indicates that the c-th channel of the b-th sample is assigned to the g-th expert. Since the argmax operation is non-differentiable, we adopt the Straight-Through Estimator (STE) [2] to decouple the discrete forward process from the continuous backward process, formulated as:

$$A = \tilde{A} + \tilde{S} - \text{sg}(\tilde{S}), \tag{7}$$

where $\text{sg}(\cdot)$ denotes the stop gradient operator. This formulation ensures that during the forward pass, the discrete matrix $\tilde{A}$ is used for routing, while during the backward pass, the gradients are passed directly to $\tilde{S}$, enabling end-to-end optimization.

Based on the routing matrix A, each input channel is exclusively assigned to its designated expert. Specifically, the set of channel indices assigned to the g-th expert E_g is formally defined as:

$$\mathcal{I}_g = \{(b,c) \mid A_{b,c,g} = 1\}, \tag{8}$$

where $b \in \{1,\dots,B\}$ and $c \in \{1,\dots,C\}$. Accordingly, we extract the input for expert g as $X_g \in \mathbb{R}^{N_g\times 1\times D\times H\times W}$, where $N_g = |\mathcal{I}_g|$, represents the total number of channels assigned to the g-th expert. Each expert is a multi-scale residual block that captures spatial context across various receptive fields. As illustrated in Figure 2 (b), for the single-channel input x_g, the expert extracts local and global features via $3\times 3\times 3$ and $5\times 5\times 5$ convolutions. The dual-branch outputs are then aggregated through channel concatenation, followed by a $1\times 1\times 1$ convolution to restore the correct channel dimensions:

$$o_g = X_g + \text{Conv}_{1\times 1\times 1}\left([\text{Conv}_{3\times 3\times 3}(X_g); \text{Conv}_{5\times 5\times 5}(X_g)]\right). \tag{9}$$

To facilitate gradient flow back to the routing matrix A, we perform element-wise multiplication between the expert outputs and the corresponding routing scores along the channel dimension:

$$y^{(g)}_{b,c} = o_{g,i} \cdot A_{b,c,g}, \tag{10}$$

where $\forall (b,c) \in \mathcal{I}_g$, $o_{g,i}$ represents the i-th channel feature ($i \in \{1,\dots,N_g\}$) output by the g-th expert. Finally, the processed channels are remapped back to their original positions in the input tensor according to the index set $\mathcal{I}_g$.

3.4 Training Loss

To mitigate the risk of expert collapse, we introduce an entropy-based load balancing loss. First, we calculate the average number of channels assigned to expert g within a batch:

$$P_g = \text{Norm}\left(\frac{1}{B}\sum_{b=1}^{B}\sum_{c=1}^{C} A_{b,c,g}\right), \tag{11}$$

where $\text{Norm}(\cdot)$ denotes the normalization operation. Unlike traditional MoE's requirement for equal data distribution per expert, we evaluate the uniformity of expert utilization by calculating the Shannon entropy of this distribution. Theoretically, the entropy reaches its maximum value of $\log(G)$ when all experts are selected with equal probability. We define the load balancing loss as the normalized divergence between the empirical entropy and its theoretical maximum, which quantifies the deviation of the current routing strategy from an ideal uniform distribution:

$$\mathcal{L}_{\text{balance}} = \frac{\log(G) + \sum_g (P_g \cdot \log(P_g + \epsilon))}{\log(G)}, \tag{12}$$

where ϵ is a small constant added for numerical stability.

Additionally, we use Dice loss and Cross-Entropy (CE) loss to jointly optimize region overlap and classification accuracy. Let $\hat{Y}$ denote the model predictions and Y represent the corresponding ground truth. The total loss function is defined as:

$$\mathcal{L} = \lambda_{\text{CE}}\,\text{CE}(\hat{Y}, Y) + \lambda_{\text{Dice}}\,\text{Dice}(\hat{Y}, Y) + \lambda_{bal}\mathcal{L}_{\text{balance}}, \tag{13}$$

where λ_i denotes the weight to the loss $\mathcal{L}_i$. To further enhance gradient propagation and training stability, deep supervision is applied at multiple decoder stages.

4 Experiments

4.1 Datasets

We construct two ankle CT datasets, namely Fuzhong and Normal, containing 107 and 155 paired scans, respectively. Among them, 11 and 15 cases are held out for testing, while the remaining data are used for five-fold cross-validation. Each CT volume is resampled to 256×256×160. All scans are fully anonymized and manually annotated by experienced radiologists using ITK-SNAP [45] into nine bone categories: Tibia, Fibula, Talus, Calcaneus, Navicular, Cuboid, and the first, second, and third Cuneiform bones. Notably, due to variations in scan range and the high cost of full labeling, only partial tibia and fibula annotations are provided in the training sets. To ensure fair evaluation, complete annotations are preserved in the testing sets to guarantee evaluation consistency.

To further validate the generalization capability of our model, we also conduct experiments on the AMOS-CT [16] dataset, a large-scale multi-organ CT segmentation dataset that includes diverse abdominal structures. Specifically, we also perform five-fold cross-validation on the training set and report the segmentation performance on the validation set. The abbreviations of the 15 anatomical structures used in Table 3 are as follows: Spl for spleen, RKid for right kidney, LKid for left kidney, Gall for gallbladder, Eso for esophagus, Liv for liver, Sto for stomach, Aor for aorta, IVC for inferior vena cava, Pan for pancreas, RAG for right adrenal gland, LAG for left adrenal gland, Duo for duodenum, Bla for bladder, and Pro/Uth for prostate/uterus.

4.2 Implementation Details

Our model is implemented based on the nnU-Net [14] framework and trained on an NVIDIA A6000 GPU. Each CT volume has a spatial size of 256 × 256 × 160, which is randomly cropped into patches of 96 × 160 × 160 during training. We adopt a five-fold cross-validation strategy across all experiments. The network is optimized using the AdamW [26] with an initial learning rate of 1e-4, a weight decay of 3e-5, and a CosineAnnealingLR scheduler. Models are trained for 1000 epochs on the ankle datasets and 2500 epochs on the AMOS dataset, with a batch size of 1. Other hyperparameters follow the default nnU-Net configurations to ensure fair comparisons.

Table 1: Quantitative comparison on the Normal dataset. Bold indicates the best.

Model	Metric	Tibia	Fibula	Talus	Calcaneus	Navicular	Cuboid	Cuneiform-1	Cuneiform-2	Cuneiform-3	Avg
3D U-Net [6]	Dice(%) ↑	60.11	75.18	95.76	**96.48**	**94.14**	92.44	92.94	91.47	92.74	87.92
	HD95(mm) ↓	32.38	32.73	1.34	**1.19**	1.39	2.25	8.47	1.42	1.36	9.17
SegResNet [29]	Dice(%) ↑	66.24	77.66	95.72	96.36	93.95	92.46	92.73	91.40	92.89	88.82
	HD95(mm) ↓	29.26	29.03	**1.32**	1.25	1.38	2.21	8.43	1.43	1.36	8.41
UMamba [28]	Dice(%) ↑	67.52	78.20	95.51	96.26	93.93	92.53	91.81	91.39	93.14	88.92
	HD95(mm) ↓	30.71	30.35	1.45	1.35	1.42	2.16	20.61	13.88	**1.29**	11.47
SegMamba [43]	Dice(%) ↑	63.27	75.89	95.64	96.43	93.92	92.47	93.20	91.50	92.72	88.34
	HD95(mm) ↓	**27.87**	29.03	1.38	1.22	1.41	2.22	8.39	1.46	1.36	8.26
LightMUNet [20]	Dice(%) ↑	63.80	74.14	95.27	96.16	92.77	91.61	92.16	88.09	90.75	87.19
	HD95(mm) ↓	31.23	31.20	1.48	1.30	1.58	2.49	1.86	1.81	1.64	8.29
MambaHoME [30]	Dice(%) ↑	58.11	76.43	**95.81**	96.44	**94.14**	**92.66**	92.88	**91.95**	93.13	87.95
	HD95(mm) ↓	32.55	29.14	1.43	1.21	1.36	**2.11**	8.40	**1.37**	1.32	8.77
3DABSeg(ours)	Dice(%) ↑	**72.14**	**81.93**	95.77	96.44	94.07	92.51	**93.98**	91.71	**93.21**	**90.20**
	HD95(mm) ↓	28.57	**25.51**	1.39	1.20	**1.34**	2.26	**1.32**	1.49	1.30	**7.15**

4.3 Evaluation Metrics

To quantitatively evaluate the segmentation performance, we adopt two metrics: the Dice coefficient and the 95th percentile Hausdorff distance (HD95). The Dice coefficient measures the volumetric overlap between the prediction and the label, reflecting overall segmentation accuracy. HD95 evaluates the spatial accuracy of predicted boundaries by computing the 95th percentile of surface distances between the prediction result and label.

4.4 Comparison with State-of-the-Arts

We compare with classic CNN-based [6, 29], Transformer-based [11, 12, 47, 42], Mamba-based [28, 43] segmentation models on the Fuzhong, Normal and AMOS-CT datasets. As shown in Table 1, on the Normal dataset, 3DABSeg achieves the highest mean Dice score of 90.20 and the lowest mean HD95 of 7.15 mm, being the only method among all compared models to break the 90 mean Dice threshold. Compared to the baseline UMamba, our method improves the Dice score by 1.28 and significantly reduces the HD95 by 4.32 mm. From the perspective of specific bone categories, 3DABSeg demonstrates an overwhelming advantage in processing challenging long bone structures.

Table 2: Quantitative comparison on the Fuzhong dataset. Bold indicates the best.

Model	Metric	Tibia	Fibula	Talus	Calcaneus	Navicular	Cuboid	Cuneiform-1	Cuneiform-2	Cuneiform-3	Avg
3D U-Net [6]	Dice(%) ↑	95.45	96.18	96.81	96.87	95.05	**95.70**	95.23	93.35	95.23	95.54
	HD95(mm) ↓	0.72	0.76	2.28	1.24	1.06	0.88	1.02	0.98	0.73	1.07
SegResNet [29]	Dice(%) ↑	95.50	96.11	**96.93**	96.88	**95.29**	95.68	95.39	93.49	95.36	95.63
	HD95(mm) ↓	0.70	0.82	**2.22**	1.23	1.06	0.87	**1.00**	0.98	0.74	1.07
UMamba [28]	Dice(%) ↑	95.30	96.22	96.86	96.94	95.22	95.67	95.32	93.37	**95.39**	95.59
	HD95(mm) ↓	0.71	0.73	2.23	**1.22**	1.02	**0.86**	1.05	0.98	0.73	**1.06**
SegMamba [43]	Dice(%) ↑	95.72	96.31	96.82	96.72	94.84	95.36	95.22	93.29	95.22	95.50
	HD95(mm) ↓	0.70	0.71	2.25	1.24	1.11	0.92	1.02	0.98	0.75	1.08
LightMUNet [20]	Dice(%) ↑	95.81	**96.48**	96.77	96.78	95.01	95.45	95.15	92.91	94.96	95.48
	HD95(mm) ↓	**0.68**	**0.67**	2.30	1.24	**0.97**	0.89	**1.00**	1.02	0.78	**1.06**
MambaHoME [30]	Dice(%) ↑	**95.84**	96.21	96.85	96.79	95.18	95.62	95.37	**93.53**	95.36	**95.64**
	HD95(mm) ↓	0.70	0.80	2.25	1.24	1.06	0.90	**1.00**	**0.95**	**0.71**	1.07
3DABSeg(ours)	Dice(%) ↑	95.41	96.42	96.87	**96.96**	95.18	95.67	**95.40**	93.52	95.37	**95.64**
	HD95(mm) ↓	0.71	0.70	2.24	1.23	1.06	**0.86**	**1.00**	0.98	0.72	**1.06**

On the Fuzhong dataset, as shown in Table 2, 3DABSeg achieves a mean Dice score of 95.64 and an HD95 of 1.06 mm, slightly surpassing strong existing models such as UMamba [28] and SegResNet [29]. The Fuzhong dataset consists of standardized standing-position scans with consistent bone orientation and minimal soft-tissue interference, making the segmentation task more straightforward and resulting in small performance differences across models. Nonetheless, 3DABSeg and MambaHoME obtain the best overall balance between Dice and HD95, accurately delineating bone boundaries while preserving anatomical consistency.

To evaluate generalization capability, we further conducted experiments on the AMOS-CT dataset, and the results are presented in Table 3. 3DABSeg achieves the highest mean Dice of 89.12 and the lowest mean HD95 of 5.60 mm, significantly outperforming all compared models. Compared with CNN-based models, the mean Dice of 3DABSeg is 5.3 higher than 3D U-Net and 9.28 higher

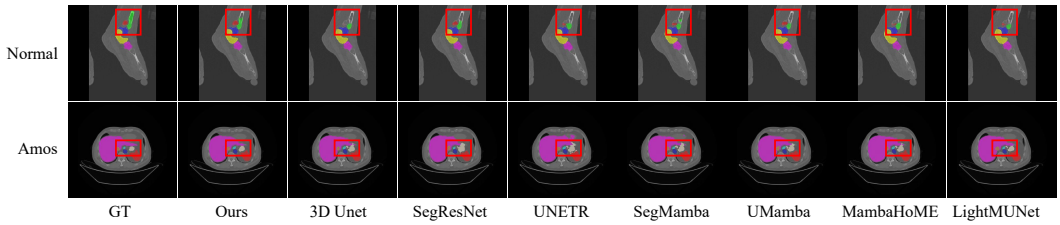

Figure 3: Qualitative comparison of segmentation results on the normal and AMOS-CT datasets.

than SegResNet. Meanwhile, when compared with Mamba-based architectures, 3DABSeg maintains a notable advantage, surpassing SegMamba and UMamba by 6.03 and 7.88 in Dice score, respectively, while substantially reducing HD95 boundary errors. Notably, 3DABSeg demonstrates an overwhelming advantage in the segmentation of highly challenging small organs, such as the Eso, RAG, LAG, and Duo, where Dice scores are improved by over 7 compared with 3D U-Net. These results provide strong evidence that 3DABSeg exhibits exceptional generalization capability in multi-organ segmentation tasks.

Table 3: Quantitative results on the AMOS-CT dataset. Bold indicates the best.

Model	Metric	Spl	RKid	LKid	Gall	Eso	Liv	Sto	Aor	IVC	Pan	RAG	LAG	Duo	Bla	Pro/Uth	**Avg**
3D U-Net [6]	Dice(%)	94.33	95.12	93.41	73.43	75.30	96.48	88.37	94.04	87.55	82.94	68.74	70.42	74.04	86.13	77.04	83.82
	HD95(mm)	3.83	**7.22**	4.86	9.24	6.60	5.80	9.64	4.67	5.05	6.16	4.60	5.56	10.00	15.15	10.26	7.24
SegResNet [29]	Dice(%)	93.54	93.21	91.65	71.32	68.52	95.78	86.99	92.40	83.97	79.13	56.96	56.49	68.61	84.30	74.65	79.84
	HD95(mm)	4.96	19.05	5.53	10.49	14.71	7.68	10.81	6.16	5.83	9.89	6.35	8.23	11.98	17.23	17.57	10.43
UNETR [12]	Dice(%)	91.95	89.86	90.02	59.73	66.51	93.88	74.92	88.58	78.33	69.56	64.51	62.13	58.21	75.97	69.50	75.58
	HD95(mm)	17.37	13.25	14.64	20.13	9.20	16.27	26.73	8.41	9.90	17.52	8.49	9.48	19.55	25.09	24.91	16.06
SegMamba [43]	Dice(%)	95.00	94.93	94.46	74.50	74.72	96.32	87.27	93.82	88.10	82.75	65.04	67.24	74.18	81.96	76.11	83.09
	HD95(mm)	6.22	10.26	8.08	14.89	8.76	8.98	13.16	4.68	9.71	9.95	9.82	10.83	13.64	25.06	13.89	11.20
UMamba [28]	Dice(%)	93.27	93.62	92.14	72.19	69.60	95.98	86.83	93.15	85.84	79.79	59.20	62.55	68.33	86.69	79.36	81.24
	HD95(mm)	5.47	10.37	5.04	9.40	8.82	7.37	13.14	6.33	5.82	6.62	6.44	7.64	11.46	11.76	12.90	8.57
MambaHoME [30]	Dice(%)	96.34	96.01	95.76	79.72	83.17	97.27	90.03	94.90	90.78	86.42	76.91	78.23	79.98	89.50	82.64	87.84
	HD95(mm)	3.43	7.59	3.23	8.12	4.73	**4.99**	7.08	4.76	4.32	6.05	3.31	4.33	**8.61**	**11.61**	9.96	6.14
LightMUNet [20]	Dice(%)	94.59	93.24	95.09	75.36	79.08	96.35	87.20	93.95	87.11	82.35	71.72	71.84	73.76	86.04	79.76	84.50
	HD95(mm)	12.56	8.42	3.43	16.07	5.95	11.25	12.25	5.86	6.28	7.41	4.78	5.62	13.40	17.41	13.90	9.64
3DABSeg (Ours)	Dice(%)	**96.91**	**96.42**	**96.17**	**82.21**	**85.61**	**97.49**	**91.45**	**95.42**	**91.60**	**87.88**	**78.12**	**79.86**	**81.94**	**90.20**	**85.48**	**89.12**
	HD95(mm)	**2.55**	7.74	**2.51**	**7.66**	**3.66**	5.06	**5.67**	**3.89**	**3.97**	**4.39**	**2.99**	**3.47**	9.12	13.80	**7.53**	**5.60**

Figure 3 presents a qualitative comparison on the Normal and AMOS datasets. For ankle bone segmentation, while existing methods struggle with the structural integrity of the tibia and fibula, 3DABSeg reconstructs more complete bone regions. On the AMOS dataset, 3DABSeg excels in delineating liver and stomach boundaries with clear anatomical margins. In contrast, competing methods produce blurred overlapping boundaries and lose critical morphological details. These results demonstrate 3DABSeg's superiority in handling complex skeletal structures and diverse abdominal organs.

4.5 Ablation study

The Number of Experts. As shown in Table 4, we explored the impact of the number of experts on model performance on the Normal dataset. The results indicate that the performance gain does not increase linearly with the number of experts. The model achieves optimal performance with three experts, resulting in a Dice score of 90.20 and an HD95 score of 7.15 mm. This configuration provides the optimal balance for ankle bone segmentation, as fewer experts hinder feature decoupling, while more experts may lead to ambiguous task allocation and increased routing complexity.

Table 4: Ablation study on the number of experts in the MFF-MoE module on the Normal dataset.

Metric	2	3	4	5
Dice (%) ↑	89.86	**90.20**	89.01	88.61
HD95 (mm) ↓	7.37	**7.15**	7.88	8.26

The Joint Effect of HKMB and MFF-MoE Modules. Compared to the baseline, individual incorporation of HKMB and MFF-MoE significantly enhances boundary precision, lowering the

HD95 to 8.36 mm and 8.22 mm. Ultimately, when HKMB and MFF-MoE work in synergy, the model achieves the best performance with a Dice of 90.20 and an HD95 of 7.15 mm. This demonstrates that the fine-grained local perception provided by HKMB is highly complementary to the global contextual modeling of MFF-MoE, obtaining the optimal segmentation results.

Table 5: Effectiveness of HKMB and MFF-MoE. Bold indicates the best.

Configuration	Dice(%) ↑	HD95(mm) ↓
baseline	88.92	11.47
HKMB only	89.27	8.36
MFF-MoE only	89.35	8.22
HKMB + MFF-MoE	**90.20**	**7.15**

Expert Specialization Figure 4 visualizes the feature activation maps of distinct experts within the MFF-MoE module. Specifically, Expert 1 primarily captures the global background context. Expert 2 focuses on modeling the spatial relationships between adjacent bone structures. In contrast, Expert 3 exhibits high-frequency responses, effectively highlighting sharp bone boundaries. These results demonstrate that by integrating the mutually exclusive hard-assignment strategy via Gumbel-Softmax, Straight-Through Estimator and entropy-based load balancing loss function, our method ensures that each expert is actively engaged and functionally complementary.

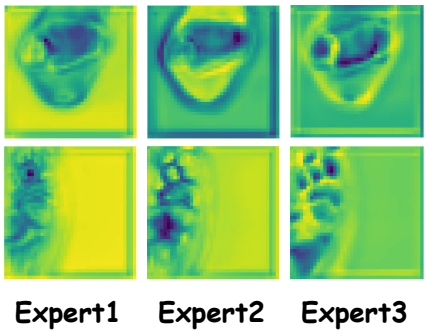

Figure 4: Visualization of the feature activation maps from experts within the MFF-MoE module.

5 Conclusion

In this paper, we propose 3DABSeg, a novel 3D framework for accurate and robust ankle bone segmentation. It is designed to address the challenges of complex anatomical structures and fine-grained bone details. Specifically, the HKMB enhances long-range spatial modeling and nonlinear channel interaction, enabling the network to progressively capture global anatomical structures and enhance subtle bone features. Meanwhile, the MFF-MoE module adaptively partitions channel features into semantically coherent groups and routes them to specialized experts, facilitating semantic decoupling and improving feature refinement. By jointly leveraging these complementary modules, 3DABSeg achieves state-of-the-art performance on several datasets, demonstrating its effectiveness in modeling intricate bone geometry and fine-grained anatomical structures.

Acknowledgments and Disclosure of Funding

This work was supported by the Shenzhen Medical Research Fund under Grant C2601012, the National Natural Science Foundation of China under Grants 62176163 and 62502320, the Natural Science Foundation of Guangdong Province under Grant 2025A1515010184, the project of Shenzhen Science and Technology Innovation Committee under Grant JCYJ20240813141424032, the Scientific Foundation for Youth Scholars of Shenzhen University, and Shenzhen Higher Education Stable Support Program General Project under Grant 20231120175215001. We also acknowledge the computational resources and support provided by the Intelligent Computing Center of Shenzhen University. The authors declare no competing interests.

References

[1] Akansh Agrawal, Akshan Agrawal, Shashwat Gupta, and Priyanka Bagade. Kan-mamba fusionnet: Redefining medical image segmentation with non-linear modeling. *arXiv preprint*, 2024.

[2] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint*, 2013.

[3] Moinak Bhattacharya, Shubham Jain, and Prateek Prasanna. Radiotransformer: a cascaded global-focal transformer for visual attention–guided disease classification. In *European Conference on Computer Vision*, pages 679–698. Springer, 2022.

[4] David Bonet, Antonio Ortega, Javier Ruiz-Hidalgo, and Sarath Shekkizhar. Channel redundancy and overlap in convolutional neural networks with channel-wise nnk graphs. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4328–4332, 2022.

[5] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint*, 2021.

[6] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016.

[7] Chenglin Fang and Kaigui Wu. Kandu-net: A dual-channel u-net with kan for medical image segmentation. *arXiv preprint*, 2024.

[8] Jin Gao, Juntu Zhao, Keyu Li, and Dequan Wang. Kan-mixer: Kolmogorov-arnold networks for gene expression prediction in plant species. In *European Conference on Computer Vision*, pages 135–150. Springer, 2024.

[9] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces, 2024.

[10] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *The International Conference on Learning Representations (ICLR)*, 2022.

[11] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI brainlesion workshop*, pages 272–284. Springer, 2021.

[12] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 574–584, 2022.

[13] Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, and Jian Wu. Unet 3+: A full-scale connected unet for medical image segmentation. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 1055–1059. Ieee, 2020.

[14] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

[15] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint*, 2016.

[16] Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *Advances in neural information processing systems*, 35:36722–36732, 2022.

[17] Andreĭ Nikolaevich Kolmogorov. *On the representation of continuous functions of several variables by superpositions of continuous functions of a smaller number of variables*. American Mathematical Society, 1961.

[18] Chenxin Li, Xinyu Liu, Wuyang Li, Cheng Wang, Hengyu Liu, Yifan Liu, Zhen Chen, and Yixuan Yuan. U-kan makes strong backbone for medical image segmentation and generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 4652–4660, 2025.

[19] Xiaomeng Li, Hao Chen, Xiaojuan Qi, Qi Dou, Chi-Wing Fu, and Pheng-Ann Heng. H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. *IEEE transactions on medical imaging*, 37(12):2663–2674, 2018.

[20] Weibin Liao, Yinghao Zhu, Xinyuan Wang, Chengwei Pan, Yasha Wang, and Liantao Ma. Lightm-unet: Mamba assists in lightweight unet for medical image segmentation. *arXiv preprint*, 2024.

[21] Ailiang Lin, Bingzhi Chen, Jiayu Xu, Zheng Zhang, Guangming Lu, and David Zhang. Ds-transunet: Dual swin transformer u-net for medical image segmentation. *IEEE Transactions on Instrumentation and Measurement*, 71:1–15, 2022.

[22] Chengliang Liu, Yuanxi Que, Wai Keung Wong, Yabo Liu, and Xiaoling Luo. Multi-view hilbert curve-based hierarchical information aggregation for incomplete multimodal alzheimer's disease diagnosis. *IEEE Transactions on Medical Imaging*, 45(7):3948–3961, 2026.

[23] Chengliang Liu, Que Yuanxi, Qihao Xu, Yabo Liu, Jie Wen, Jinghua Wang, and Xiaoling Luo. Hierarchical information aggregation for incomplete multimodal alzheimer's disease diagnosis. *Advances in Neural Information Processing Systems*, 38:53848–53869, 2026.

[24] Jiarun Liu, Hao Yang, Hong-Yu Zhou, Yan Xi, Lequan Yu, Cheng Li, Yong Liang, Guangming Shi, Yizhou Yu, Shaoting Zhang, et al. Swin-umamba: Mamba-based unet with imagenet-based pretraining. In *International conference on medical image computing and computer-assisted intervention*, pages 615–625. Springer, 2024.

[25] Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljačić, Thomas Y Hou, and Max Tegmark. Kan: Kolmogorov-arnold networks. *arXiv preprint*, 2024.

[26] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint*, 2017.

[27] Zichen Luo, Xinshan Zhu, Lan Zhang, and Biao Sun. Rethinking u-net: Task-adaptive mixture of skip connections for enhanced medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 5874–5882, 2025.

[28] Jun Ma, Feifei Li, and Bo Wang. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint*, 2024.

[29] Andriy Myronenko. 3d mri brain tumor segmentation using autoencoder regularization. In *International MICCAI brainlesion workshop*, pages 311–320. Springer, 2018.

[30] Szymon Płotka, Gizem Mert, Maciej Chrabaszcz, Ewa Szczurek, and Arkadiusz Sitek. Mamba goes home: Hierarchical soft mixture-of-experts for 3d medical image segmentation. *arXiv preprint*, 2025.

[31] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[32] Jiacheng Ruan, Jincheng Li, and Suncheng Xiang. Vm-unet: Vision mamba unet for medical image segmentation. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2024.

[33] Abdelrahman Shaker, Muhammad Maaz, Hanoona Rasheed, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Unetr++: delving into efficient and accurate 3d medical image segmentation. *IEEE Transactions on Medical Imaging*, 43(9):3377–3390, 2024.

[34] Abdelrahman Shaker, Syed Talal Wasim, Salman Khan, Juergen Gall, and Fahad Shahbaz Khan. Group-mamba: Efficient group-based visual state space model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14912–14922, 2025.

[35] Shengbo Tan, Rundong Xue, Shipeng Luo, Zeyu Zhang, Xinran Wang, Lei Zhang, Daji Ergu, Zhang Yi, Yang Zhao, and Ying Cai. Segkan: High-resolution medical image segmentation with long-distance dependencies. *arXiv preprint*, 2024.

[36] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, 34:24261–24272, 2021.

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[38] Risheng Wang, Tao Lei, Ruixia Cui, Bingtao Zhang, Hongying Meng, and Asoke K Nandi. Medical image segmentation using deep learning: A survey. *IET image processing*, 16(5):1243–1267, 2022.

[39] Wenxuan Wang, Chen Chen, Meng Ding, Hong Yu, Sen Zha, and Jiangyun Li. Transbts: Multimodal brain tumor segmentation using transformer. In *International conference on medical image computing and computer-assisted intervention*, pages 109–119. Springer, 2021.

[40] Boqian Wu, Qiao Xiao, Shiwei Liu, Lu Yin, Mykola Pechenizkiy, Decebal Constantin Mocanu, Maurice Keulen, and Elena Mocanu. E2enet: Dynamic sparse feature fusion for accurate and efficient 3d medical image segmentation. *Advances in Neural Information Processing Systems*, 37:118483–118512, 2024.

[41] Yanlin Wu, Tao Li, Zhihong Wang, Hong Kang, and Along He. Transukan: computing-efficient hybrid kan-transformer for enhanced medical image segmentation. *arXiv preprint*, 2024.

[42] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

[43] Zhaohu Xing, Tian Ye, Yijun Yang, Guang Liu, and Lei Zhu. SegMamba: Long-range Sequential Modeling Mamba For 3D Medical Image Segmentation . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15008. Springer Nature Switzerland, October 2024.

[44] Guoping Xu, Xuan Zhang, Xinwei He, and Xinglong Wu. Levit-unet: Make faster encoders with transformer for medical image segmentation. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 42–53. Springer, 2023.

[45] Paul A. Yushkevich, Joseph Piven, Heather Cody Hazlett, Rachel Gimpel Smith, Sean Ho, James C. Gee, and Guido Gerig. User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *Neuroimage*, 31(3):1116–1128, 2006.

[46] Yibo Zhang. Km-unet kan mamba unet for medical image segmentation. *arXiv preprint*, 2025.

[47] Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Xiaoguang Han, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnformer: Volumetric medical image segmentation via a 3d transformer. *IEEE transactions on image processing*, 32:4036–4045, 2023.

[48] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *International workshop on deep learning in medical image analysis*, pages 3–11. Springer, 2018.

A Datasets

As shown in Figure 5, the Fuzhong dataset captures CT scans acquired in a standing posture under natural load-bearing conditions, whereas the Normal dataset consists of scans collected in a supine position. Consequently, the ankle poses vary widely across datasets, and the Normal scans occasionally contain visible artifacts from imaging equipment, which are discernible to human observers but challenging for models to distinguish.

In practical clinical scenarios, the captured range of the tibia and fibula varies due to differences in the starting position and posture of each patient during CT imaging. Given the prohibitive cost of full manual annotation for these bones, our training set provides only partial masks for the tibia and fibula. This setup is intended to simulate real-world data incompleteness, challenging the model's ability to learn from local segments and generalize to the complete anatomical structure. However, to ensure evaluation fairness, we preserve complete annotations for the tibia and fibula in the testing set.

The AMOS-CT is a large-scale, clinical dataset designed for multi-organ segmentation, providing high-quality voxel-level annotations for 15 abdominal organs across diverse CT scans. To ensure clarity and conciseness in our experimental results tables, we utilize standardized abbreviations for these anatomical structures. The mapping from abbreviation to full name is as follows: Spl for spleen, RKid for right kidney, LKid for left kidney, Gall for gallbladder, Eso for esophagus, Liv for liver, Sto for stomach, Aor for aorta, IVC for inferior vena cava, Pan for pancreas, RAG for right adrenal gland, LAG for left adrenal gland, Duo for duodenum, Bla for bladder, and Pro/Uth for prostate/uterus.

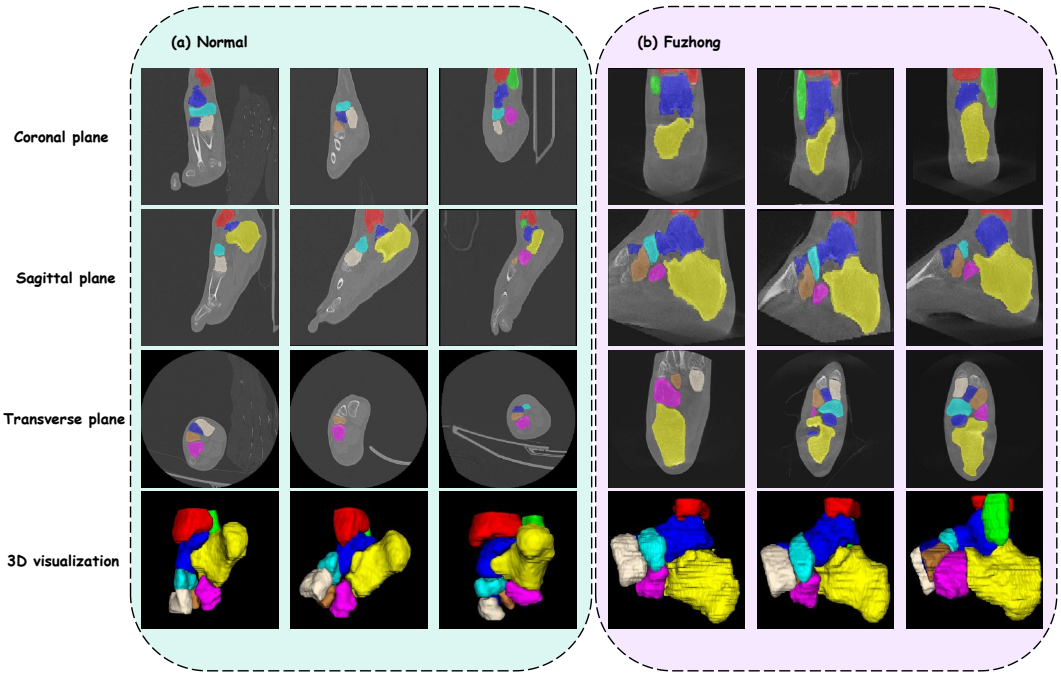

Figure 5: Visualization of the Fuzhong and Normal datasets.

B Related Work

B.1 Medical Image Segmentation

Since its introduction in 2015 by Ronneberger et al. [31], the U-Net architecture has rapidly become the backbone of medical image segmentation due to its strong performance and symmetric encoder–decoder design. Since CT and MRI are inherently 3D, Çiçek et al. [6] extended it to 3D U-Net, enabling networks to directly process volumetric data. Building on this foundation, a series of variants have emerged, such as UNet++ [48], UNet3+ [13], nnU-Net [14], and e2eNet [40], which further refine the architecture by enhancing feature aggregation. However, the limited receptive field

of CNN-based models constrains their ability to capture global contextual information. To address this problem, researchers [12, 33, 5, 44, 21, 39] integrated Transformer [37] into CNN frameworks, enhancing long-range modeling while preserving local feature extraction. Although effective, they often rely on large-scale annotated datasets and intensive computational resources, limiting their applicability in real-world scenarios.

In recent years, State Space Models [10] have emerged as a promising approach for medical image segmentation due to their favorable trade-off between efficiency and performance. U-Mamba [28] combined the local feature extraction capability of convolutions with the global modeling ability of SSMs, thereby integrating local and global representations. SegMamba [43] introduced the Tri-orientated Mamba module to model spatial dependencies through multi-directional scanning, alleviating directional sensitivity in 3D data. VM-UNet [32] proposed the VSS module and an asymmetric encoder-decoder architecture to enhance contextual representations. Furthermore, other approaches such as Swin-UMamba [24] and LightM-UNet [20] explored the applicability of Mamba in medical segmentation through windowed mechanisms and lightweight design. Although Mamba demonstrates strong capability in modeling long-range dependencies, its inherently linear formulation limits its ability to capture complex nonlinear patterns and fine-grained anatomical details, which are crucial for accurate medical image segmentation.

B.2 Kolmogorov-Arnold Networks

The Kolmogorov–Arnold theorem states that any continuous multivariate function can be represented as a weighted sum of finitely many continuous univariate functions, providing a theoretical foundation for universal approximation in neural networks [17, 25]. Unlike traditional activation functions, such as ReLU or Sigmoid, Kolmogorov–Arnold Networks (KANs) [25] treat the activation function as a learnable mapping, improving interpretability and flexibility while maintaining universal approximation capability.

Recent studies have expanded the application of KANs to visual tasks by integrating them with various mainstream architectures to enhance nonlinear feature representation. U-KAN [18], SegKAN [35] and KANDU-Net [7] incorporate KAN into the U-Net framework, introducing learnable functional mappings, Fourier-based KAN convolutions, and KAN–convolution dual-channel designs to strengthen feature fusion and nonlinear expressivity. KM-UNet [46] and KAN-Mamba FusionNet [1] leverage KAN's continuous functional approximation and Mamba's efficient long-range dependency modeling to achieve effective coordination between local and global representations. Moreover, TransUKAN [41] and KAN-Mixer [8] integrate KAN with Transformer and MLP-Mixer [36] architectures respectively, enhancing adaptive multi-scale modeling. Overall, these studies highlight KAN's flexibility and interpretability, as well as its strong adaptability and performance advantages across diverse visual architectures.

C Additional quantitative analysis on ankle datasets

Table 6 presents a quantitative comparison of more methods on Fuzhong dataset. Our proposed 3DABSeg outperforms UNETR [12], Swin UNETR [11], nnFormer [47] and SegFormer [42]. Specifically, 3DABSeg achieves the highest mean Dice of 95.64, with superior performance in key anatomical structures such as the Calcaneus and Cuboid. Regarding the HD95 metric, 3DABSeg demonstrates higher boundary precision, attaining the best results on the Cuboid and Cuneiform-3. Notably, compared to the Transformer-based SegFormer, 3DABSeg delivers a significant performance gain of 1.84 in terms of Dice, showcasing its robustness in capturing complex ankle bone morphologies.

As shown in Table 7, on Normal dataset, although nnFormer achieves the second-best results in boundary precision on normal dataset, its Dice score is 5 lower than that of 3DABSeg. On the Tibia and Fibula, our method achieves Dice values of 72.14 and 81.93 respectively, significantly outperforming traditional models such as 3D U-Net and the Transformer-based Swin UNETR. From the perspective of specific bone categories, 3DABSeg demonstrates an overwhelming advantage in processing challenging long bone structures.

Table 6: More quantitative comparison on the Fuzhong datasets. Bold indicates the best.

Model	Metric	Tibia	Fibula	Talus	Calcaneus	Navicular	Cuboid	Cuneiform-1	Cuneiform-2	Cuneiform-3	Mean
UNETR [12]	Dice(%) ↑	94.99	96.22	96.77	96.76	95.10	95.37	94.93	93.23	95.04	95.38
	HD95(mm) ↓	0.80	0.74	2.26	1.23	0.98	0.91	**1.00**	0.99	0.77	1.08
Swin UNETR [11]	Dice(%) ↑	**95.43**	96.04	**96.87**	96.89	**95.21**	95.58	95.26	93.43	95.23	95.55
	HD95(mm) ↓	0.73	0.88	**2.23**	**1.21**	1.09	0.88	1.09	**0.96**	0.74	1.09
nnFormer [47]	Dice(%) ↑	94.57	95.85	96.34	96.63	94.57	94.64	93.65	90.62	93.47	94.48
	HD95(mm) ↓	0.78	0.73	2.35	1.28	**0.92**	1.03	1.33	1.25	0.96	1.18
SegFormer [42]	Dice(%) ↑	94.06	95.37	95.71	96.20	93.81	93.78	93.30	90.23	91.78	93.80
	HD95(mm) ↓	0.81	0.85	2.48	1.43	1.05	1.16	1.37	1.29	1.24	1.30
3DABSeg(ours)	Dice(%) ↑	95.41	**96.42**	**96.87**	**96.96**	95.18	**95.67**	**95.40**	**93.52**	**95.37**	**95.64**
	HD95(mm) ↓	**0.71**	**0.70**	2.24	1.23	1.06	**0.86**	**1.00**	0.98	**0.72**	**1.06**

Table 7: More quantitative comparison on the Normal dataset. Bold indicates the best.

Model	Metric	Tibia	Fibula	Talus	Calcaneus	Navicular	Cuboid	Cuneiform-1	Cuneiform-2	Cuneiform-3	Mean
UNETR [12]	Dice(%) ↑	65.52	74.28	94.13	95.43	92.46	89.14	91.25	85.26	85.02	85.83
	HD95(mm) ↓	28.11	30.34	1.96	1.75	2.05	3.10	13.30	2.22	2.36	9.47
Swin UNETR [11]	Dice(%) ↑	68.87	78.24	95.68	96.35	**94.19**	**92.59**	92.67	91.11	92.84	89.17
	HD95(mm) ↓	27.54	27.57	1.44	1.27	1.38	**2.18**	8.47	**1.45**	1.32	8.07
nnFormer [47]	Dice(%) ↑	65.21	79.04	92.93	94.98	90.58	88.76	89.55	81.50	84.81	85.26
	HD95(mm) ↓	28.70	27.54	2.01	1.68	1.96	2.73	1.92	2.24	2.44	7.91
SegFormer [42]	Dice(%) ↑	69.68	79.79	88.84	92.58	83.53	80.89	82.66	71.04	76.27	80.59
	HD95(mm) ↓	**26.59**	**24.30**	3.24	3.08	2.79	4.39	3.10	3.55	3.28	8.26
3DABSeg(ours)	Dice(%) ↑	**72.14**	**81.93**	**95.77**	**96.44**	94.07	92.51	**93.98**	**91.71**	**93.21**	**90.20**
	HD95(mm) ↓	28.57	25.51	**1.39**	**1.20**	**1.34**	2.26	**1.32**	1.49	**1.30**	**7.15**

D Computational Efficiency

Table 8 presents the parameter counts and single-sample inference times for all comparative methods. Although models such as SegFormer [42] and LightMUNet [20] offer certain advantages in terms of parameter scale and inference speed, their average Dice scores and HD95 metrics are significantly inferior to those of 3DABSeg when processing the anatomically complex Normal dataset. In contrast, 3DABSeg, with a moderate parameter size of 66.94 M, achieves the SOTA Dice score of 89.12 on the AMOS dataset, substantially outperforming traditional CNNs and Transformer-based architectures. Notably, a comparison with MambaHome [30], which also integrates Mamba modules and the MoE, reveals that 3DABSeg requires only approximately 39% of the parameters while delivering a nearly 28% improvement in inference speed under similar core components. Furthermore, 3DABSeg achieves superior boundary prediction accuracy, as evidenced by the HD95 metric. These results provide robust evidence that 3DABSeg achieves a highly efficient synergy between Mamba modules and the Mixture of Experts model.

Table 8: Comparison of model complexity and efficiency. Parameters are in Millions (M), and Inference Time is in seconds (s). Bold indicates the best performance in each category.

Model	Parameters (M)	Inference Time (s)
3D U-Net	27.64	12.23
SegResNet	18.80	14.36
UNETR	93.08	**2.21**
Swin UNETR	62.19	26.92
nnFormer	149.51	6.20
SegFormer	**4.49**	6.01
U-Mamba	42.88	7.32
SegMamba	66.85	7.02
LightMUNet	5.02	12.65
MambaHome	170.13	29.01
3DABSeg (Ours)	66.94	20.94

E Ablation study on the weight of entropy-based load balancing loss

Table 9 illustrates the impact of entropy-based load balancing loss weight on the segmentation performance. Experimental results indicate that this weight is crucial for balancing the diversity of the expert networks and the overall segmentation accuracy. When the weight is set to 0.01, the model achieves optimal performance with a Dice score of 90.20 and an HD95 of 7.15 mm. Further observation reveals that weights that are either too high at 0.1 or too low at 0.001 lead to a decline in performance. This trend suggests that an appropriate diversity constraint effectively prevents expert collapse by encouraging experts to explore distinct semantic subspaces.

Table 9: Ablation study on the weight of entropy-based load balancing loss. Bold indicates the best results.

Loss Weight	Dice(%) ↑	HD95(mm) ↓
$\lambda_{bal} = 0$	89.12	8.20
$\lambda_{bal} = 0.1$	89.02	7.80
$\lambda_{bal} = 0.01$	**90.20**	**7.15**
$\lambda_{bal} = 0.001$	88.79	8.67

Table 10 presents the expert load distribution across 64 channels under various entropy-based load balancing loss weights λ_{bal}. Without the loss, the model suffers from severe 'expert collapse', where all channels are assigned to a single expert. While weights of 0.1 and 0.001 partially activate the multi-expert mechanism, they still result in an imbalanced load distribution. In contrast, when $\lambda_{bal} = 0.01$, the model achieves a near-uniform distribution. This balanced utilization ensures that each expert learns specialized feature representations, thereby maximizing the overall representational capacity of the MFF-MoE module.

Table 10: Expert load distribution across 64 channels under different entropy-based load balancing loss weights λ_{bal}.

Configuration	Expert 1	Expert 2	Expert 3
$\lambda_{bal} = 0$	64	0	0
$\lambda_{bal} = 0.1$	5	59	0
$\lambda_{bal} = 0.001$	36	28	0
Ours ($\lambda_{bal} = 0.01$)	**20**	**23**	**21**

F The Application Stage of the MFF-MoE Module

To examine the influence of applying the MFF-MoE module at different stages, we conduct an ablation study as shown in Table 11. The best performance is achieved when the module is applied at the intermediate stages (2, 3). In contrast, applying it to deeper stages results in a slight performance drop, which is likely attributable to the reduced spatial resolution of deeper features, which limits the effectiveness of expert grouping.

Table 11: Ablation study on the application stage of the MFF-MoE Module. Bold indicates the best.

Metric	(2)	(2,3)	(1,2,3)	(0,1,2,3)
Dice (%) ↑	88.73	**90.20**	88.64	88.80
HD95 (mm) ↓	8.07	**7.15**	8.06	8.75

G The spline order in the KAN module

As shown in Table 12, the spline order notably affects model performance. The best results are obtained with a cubic spline, achieving the highest Dice and the lowest HD95. Lower-order splines offer limited representational flexibility, whereas higher orders may introduce overfitting and unstable

boundaries. Overall, the cubic spline achieves the best balance between nonlinear expressiveness and stability.

Table 12: Ablation study of spline order in the KAN module. Bold indicates the best.

Metric	1	2	3	4
Dice(%) ↑	89.54	88.87	**90.20**	89.16
HD95(mm) ↓	7.44	9.47	**7.15**	8.40

H Ablation study on sampling strategies

Table 13 presents the ablation results regarding the multi-scale sampling strategies within the MFF-MoE module. The experiments compare configurations utilizing a single adaptive average pooling scale among 1, 4, or 8 with the proposed combined strategy. The results indicate that a single sampling scale is insufficient to balance global context and local structural details. Specifically, the only 4-scale sampling outperforms the extremely low-resolution only 1-scale global sampling, with the Dice score increasing from 88.97 to 89.93, highlighting the importance of finer spatial feature representation. However, further increasing the sampling scale to 8 leads to a performance decline, as evidenced by the HD95 rising to 8.28 mm which suggests that excessive spatial resolution may introduce redundant local noise and complicate feature aggregation.

In contrast, the proposed strategy achieves the best segmentation performance by integrating a multi-scale combination of 1, 4, and 8, reaching a Dice score of 90.20 and an HD95 of 7.15 mm. These findings provide robust evidence that by fusing sampling branches with varying receptive fields, the model can simultaneously capture low-resolution global semantics and high-resolution anatomical details, thereby enabling more precise feature extraction within a complex multi-scale feature space.

Table 13: Ablation study on multi-scale sampling strategies within MFF-MoE. Bold indicates the best.

Configuration	Dice(%) ↑	HD95(mm) ↓
Only 1 AvgPool	88.97	7.66
Only 4 AvgPool	89.93	7.17
Only 8 AvgPool	89.21	8.28
Ours	**90.20**	**7.15**

I Limitations

While 3DABSeg achieves SOTA performance on the ankle and abdomen CT datasets, several limitations remain to be addressed in future research.

First, a significant challenge arises from label incompleteness in the current training set. Certain anatomical structures, such as the tibia and fibula, are partially annotated in some training samples, whereas the test set requires complete and precise segmentation. This discrepancy between incomplete training supervision and complete inference requirements poses a challenge for the model to maintain semantic consistency across global structures.

Second, the inductive bias of scanning strategies in our Hybrid KAN-Mamba Block (HKMB) warrants further investigation. Although Mamba's linear-complexity sequence modeling effectively captures long-range dependencies, the choice of 3D-to-1D serialization trajectories may introduce orientation-specific sensitivity. Whether different scanning orders or multi-directional fusion can further enhance the robustness of complex anatomical modeling remains an open question.

J Broader Impact

Our work introduces two ankle CT datasets and a novel 3DABSeg network, providing significant clinical value by assisting surgeons in precise preoperative planning and alleviating the intensive workload of manual annotation.

However, certain ethical and technical risks warrant consideration. First, regarding privacy risks, although all imaging data in our dataset have undergone rigorous de-identification and anonymization, the large-scale sharing of medical images in an open-science context necessitates stringent data governance.

Second, the generalization across diverse imaging protocols remains a critical consideration. While 3DABSeg achieves SOTA performance on above datasets, its robustness across CT scanners from different manufacturers or with varying acquisition parameters requires further validation. Users should be mindful of potential performance fluctuations when deploying the model in heterogeneous clinical environments.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We have claimed that our proposed 3D Ankle Bone Segmentation network (3DABSeg) in the abstract and introduction.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We provide a comprehensive discussion of the limitations in Appendix I. Specifically, we reflect on the challenges posed by partial annotations of tibia and fibula in the training set, and the potential influence of Mamba's scanning trajectories on 3D spatial consistency.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: We do not introduce new mathematical theorems or formal theoretical proofs that require separate assumptions and verification.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide implementation details in Implementation Details to ensure reproducibility.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The source code is publicly available at `https://github.com/Yikol/3DABSeg`. However, the two ankle CT datasets cannot be publicly released due to patient privacy restrictions. The AMOS dataset used in our experiments is publicly available.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide all training and testing details in Implementation Details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We do not report error bars or statistical significance tests in our tables.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All experiments were conducted on NVIDIA A6000 GPUs, and we have reported the number of parameters and the average inference time per sample for each method in Appendix D.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We provide a discussion of the potential positive and negative societal impacts of our work in the Broader Impact section of Appendix J.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: All CT scans were de-identified to protect patient privacy. The two ankle CT datasets are not publicly released due to patient privacy restrictions.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have properly cited the original papers for all existing assets used, and we strictly adhere to their respective licenses and terms of use for academic research.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The ankle CT datasets are described in the paper and supplementary material, but cannot be publicly released due to patient privacy restrictions. The 3DABSeg source code is publicly available at `https://github.com/Yikol/3DABSeg`.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Our research involves retrospective analysis of anonymized clinical CT scans. All data were de-identified to protect patient privacy in accordance with ethical standards. No crowdsourcing platforms were used for data annotation.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

Justification: This study is a retrospective analysis of fully de-identified and anonymized clinical CT scans obtained from hospital archives. The data were collected during routine clinical care and stripped of all protected health information before being accessed for this research. As the study involves no intervention and poses no risk to participants, it follows the standard ethical practices for using anonymized retrospective data.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: Our method and results do not involve the use of LLMs.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.