Hybrid Transformer-Mamba Architecture for Weakly Supervised Volumetric Medical Segmentation

Yiheng Lyu[1,2] (ID)   YIHENG.LYU@UWA.EDU.AU

Lian Xu[1] (ID)   LIAN.XU@UWA.EDU.AU

Mohammed Bennamoun[1] (ID)   MOHAMMED.BENNAMOUN@UWA.EDU.AU

Farid Boussaid[3] (ID)   FARID.BOUSSAID@UWA.EDU.AU

Coen Arrow[2,4] (ID)   COEN.ARROW@RESEARCH.UWA.EDU.AU

Girish Dwivedi[2,4,5,6] (ID)   GIRISH.DWIVEDI@PERKINS.UWA.EDU.AU

[1] *School of Physics, Mathematics and Computing, University of Western Australia, Perth, Australia*

[2] *Harry Perkins Institute of Medical Research, Perth, Australia*

[3] *School of Engineering, University of Western Australia, Perth, Australia*

[4] *Medical School, University of Western Australia, Perth, Australia*

[5] *Fiona Stanley Hospital, Perth, Australia*

[6] *Victor Chang Cardiac Research Institute, Sydney, Australia*

Abstract

Weakly supervised semantic segmentation offers a label-efficient solution to train segmentation models for volumetric medical imaging. However, existing approaches often rely on 2D encoders that neglect the inherent volumetric nature of the data. We propose TranSamba, a hybrid Transformer-Mamba architecture designed to capture 3D context for weakly supervised volumetric medical segmentation. TranSamba augments a standard Vision Transformer backbone with Cross-Plane Mamba blocks, which leverage the linear complexity of state space models for efficient information exchange across neighboring slices. The information exchange enhances the pairwise self-attention within slices computed by the Transformer blocks, directly contributing to the attention maps for object localization. TranSamba achieves effective volumetric modeling with time complexity that scales linearly with the input volume depth and maintains constant memory usage for batch processing. Extensive experiments on three datasets demonstrate that TranSamba establishes new state-of-the-art performance, consistently outperforming existing methods across diverse modalities and pathologies. Our source code and trained models are openly accessible at: `https://github.com/YihengLyu/TranSamba`.

Keywords: Computed tomography, magnetic resonance imaging, Mamba, medical segmentation, state space model, weakly supervised learning.

1. Introduction

Volumetric medical imaging including computed tomography (CT) and magnetic resonance imaging (MRI) is acquired by volumetric reconstruction of 3D space, enabling non-invasive anatomical examination. Image segmentation is important for image-guided intervention and can be automated with deep learning (Ronneberger et al., 2015). However, model training generally relies on pixel-level labels, which are labor-intensive to obtain at scale. Weakly supervised semantic segmentation (WSSS) offers a label-efficient alternative by training segmentation models with weak labels that are less expensive to obtain. A common WSSS

strategy uses class activation maps (CAMs) (Zhou et al., 2016), which are extracted from trained classification models and encode the localization information of discriminative regions that drive model decision-making. WSSS is crucial for volumetric medical imaging, as obtaining voxel-level labels requires significant expertise and is exceptionally costly. Meanwhile, obtaining slice-level labels for an anatomical object can be simplified to identifying only its first and last slices, which is highly cost-effective.

However, CT and MRI have domain-specific characteristics that make CAM-based object localization challenging. Research in weakly supervised medical segmentation[1] has sought to address these challenges, for example, through multi-scale fusion to improve low-resolution CAMs (Ma et al., 2020; Chen et al., 2023b; Fu et al., 2023), using self-supervision to bridge the slice-to-voxel gap (Tang et al., 2021a; Kuang et al., 2023; Lyu et al., 2024), and leveraging domain-prior constraints (Chen et al., 2022; Li et al., 2022; Chen et al., 2023a). Nevertheless, the volumetric nature, which makes CT and MRI fundamentally distinct from natural images, remains largely underexplored in WSSS based on slice-level labels.

With voxel-level labels, volumetric modeling can be achieved with 3D encoders (Milletari et al., 2016) in a fully supervised manner. However, two factors favor 2D encoders in weakly supervised settings. First, only 2D encoders can make predictions from individual slices that align directly with slice-level labels. Second, 2D encoders like the Vision Transformer (ViT) (Dosovitskiy, 2020) show strong transferability when pre-trained. Specifically, class-to-patch (C2P) attention, the pairwise self-attention (SA) (Vaswani et al., 2017) between class and patch tokens, has proven effective for object localization in WSSS (Xu et al., 2022, 2024). Here, we term the SA between the class and patch tokens from a single 2D image or slice as **in-plane** SA. Only in-plane SA can be learned using a 2D ViT encoder. The interactions between tokens across multiple neighboring planes, which we term **cross-plane** interactions, complement in-plane SA. To achieve volumetric modeling with a 2D encoder, cross-plane interactions must be integrated with in-plane SA.

However, the quadratic-time computation of SA makes it inefficient for cross-plane modeling. The time complexity of cross-plane SA across N planes is N^2 times its in-plane counterpart. With limited computational resources, computing cross-plane SA can easily lead to out-of-memory issues. Built on a structured state space model (SSM), the Mamba architecture (Gu and Dao, 2024) achieves linear-time scaling with respect to sequence length by incorporating a selective mechanism that dynamically filters input tokens based on their relevance. Because of its computational efficiency, Mamba has been introduced to computer vision (Zhu et al., 2024; Liu et al., 2024c). In medical imaging, the applications of Mamba[2] have proven successful for medical segmentation (Ma et al., 2024; Xing et al., 2024), including under weakly supervised settings (Fan et al., 2024). The balanced efficiency and effectiveness of Mamba make it a promising solution for cross-plane modeling.

To this end, we propose TranSamba, a hybrid **TranS**former-M**amba** architecture for cross-plane modeling to enhance weakly supervised volumetric medical segmentation. TranSamba is built on a ViT encoder and features an intuitive design: each layer of the encoder is augmented with a Cross-Plane Mamba (CPM) block, placed in series with the Transformer block. The CPM blocks leverage SSM for efficient cross-plane modeling, while the Transformer blocks learn in-plane SA. Cross-plane modeling improves the quality of in-plane C2P

---

1. Related work on weakly supervised medical segmentation is presented in Appendix A.1
2. Related work on Mamba in medical imaging is presented in Appendix A.2

attention by enriching the information encoded in each token. With the simple addition of a CPM block per layer, TranSamba achieves accurate object localization via single-stage training. **The main contributions of this paper can be summarized as follows: 1.** We propose TranSamba, a hybrid architecture that combines Transformer for powerful in-plane SA with Mamba for efficient cross-plane modeling via SSM. **2.** The time complexity of TranSamba scales linearly with the number of planes; for batch processing, the memory usage remains constant with respect to the number of planes. **3.** TranSamba establishes new state-of-the-art (SOTA) performance on three datasets covering distinct pathologies and modalities, underscoring the effectiveness of cross-plane modeling for weakly supervised volumetric medical segmentation.

2. Methodology

2.1. Overview

An overview of TranSamba is shown in Figure 1. The encoder comprises L hybrid layers; each layer consists of a CPM block (Section 2.2) in series with a vanilla Transformer block, with a residual connection applied after the CPM block.

An input volume comprises N planes; each plane is transformed into a sequence of tokens $\mathbf{T}_0^{'} \in \mathbb{R}^{1 \times (1+M) \times D}$ using the ViT image-to-sequence transformation, where M denotes the number of patch tokens per sequence and D denotes the embedding dimension.

In the l-th layer of the encoder, the input to the CPM block is obtained by stacking the output sequences from the $(l-1)$-th layer:

$$\mathbf{V}_l = [[\mathbf{T}_{l-1,1}^{'}], \ldots, [\mathbf{T}_{l-1,N}^{'}]] \tag{1}$$

The CPM block performs cross-plane modeling via SSM:

$$\mathbf{V}_l^{'} = \text{CPM}(\mathbf{V}_l) + \mathbf{V}_l \tag{2}$$

The CPM output is reshaped back into N sequences:

$$\mathbf{T}_{l,n} = \mathbf{V}_l^{'}[n,:,:], \quad n = 1, \ldots, N \tag{3}$$

The Transformer block computes in-plane SA via parallelization:

$$\mathbf{T}_{l,n}^{'}, \mathbf{A}_{l,n} = \text{Transformer}(\mathbf{T}_{l,n}), \quad n = 1, \ldots, N \tag{4}$$

where $\mathbf{A}_{l,n} \in \mathbb{R}^{(1+M) \times (1+M)}$ denotes the pairwise SA.

During training, the output sequence $\mathbf{T}_L^{'}$ consists of a class token $\mathbf{T}_L^{class} \in \mathbb{R}^{1 \times D}$ and M patch tokens $\mathbf{T}_L^{patch} \in \mathbb{R}^{M \times D}$. $\mathbf{T}_L^{class}$ and $\mathbf{T}_L^{patch}$ are separately fed into two parallel class prediction branches of the classification head. The first branch processes the class token using a global average pooling (GAP) layer:

$$\hat{y}^{class} = \text{GAP}(\mathbf{T}_L^{class}) \tag{5}$$

The second branch processes the patch tokens using a convolutional layer followed by a global weighted ranking pooling (GWRP) (Kolesnikov and Lampert, 2016) layer:

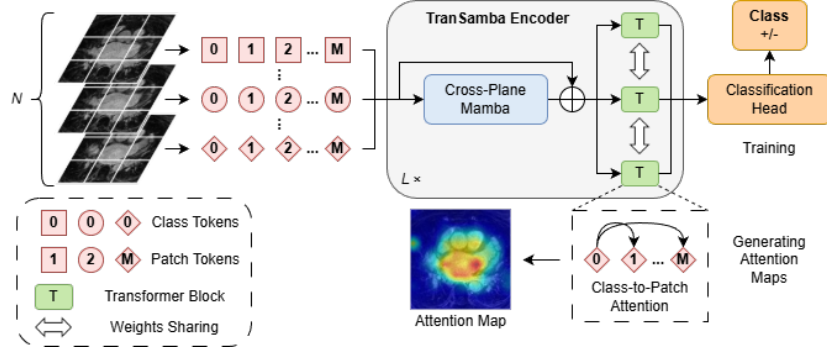$$\hat{y}^{patch} = \text{GWRP}(\text{Conv}(\mathbf{T}_L^{patch})) \tag{6}$$

Figure 1: An overview of TranSamba. The encoder consists of L hybrid layers of CPM and Transformer blocks. An input volume consists of N planes, each transformed into a sequence of $1 + M$ tokens; different shapes represent tokens from different planes. Upon entering a CPM block, the N sequences are reshaped into a single sequence and fed into a Mamba block. After cross-plane modeling, the Mamba output is reshaped back to N sequences, and the Transformer block computes in-plane SA via parallelization. During training, class scores are computed from the encoder output. For inference, attention maps are generated from the class-to-patch attention of the L Transformer blocks.

The binary cross-entropy (BCE) loss between the class scores and the slice-level label is computed as

$$\mathcal{L} = \text{BCE}(\hat{y}^{class}, y) + \text{BCE}(\hat{y}^{patch}, y) \tag{7}$$

For inference, attention maps are generated by extracting and summing C2P attention maps from all L Transformer blocks

$$\mathbf{A}^{\text{C2P}} = \sum_{l=1}^{L} \mathbf{A}_l[1, 2 :] \tag{8}$$

then reshaping and interpolating them to the original spatial resolution.

2.2. Cross-Plane Mamba

The proposed CPM block, illustrated in Figure 2, comprises a vanilla Mamba (Gu and Dao, 2024) block and two reshaping functions. It accepts an input $\mathbf{V} \in \mathbb{R}^{N \times (1+M) \times D}$ comprising N sequences. Each sequence contains one class token, serving as the global representation of its corresponding plane, and M patch tokens representing local patches. Cross-plane modeling is achieved by learning interactions between the $M \times N$ patch tokens via the Mamba block.

The N class tokens are first separated from the $N \times M$ patch tokens and excluded from the Mamba block. The $N \times M$ patch tokens are reshaped into a single sequence $\mathbf{x} \in \mathbb{R}^{(M \times N) \times D}$. Mamba employs a selection mechanism for data-dependent, non-pairwise
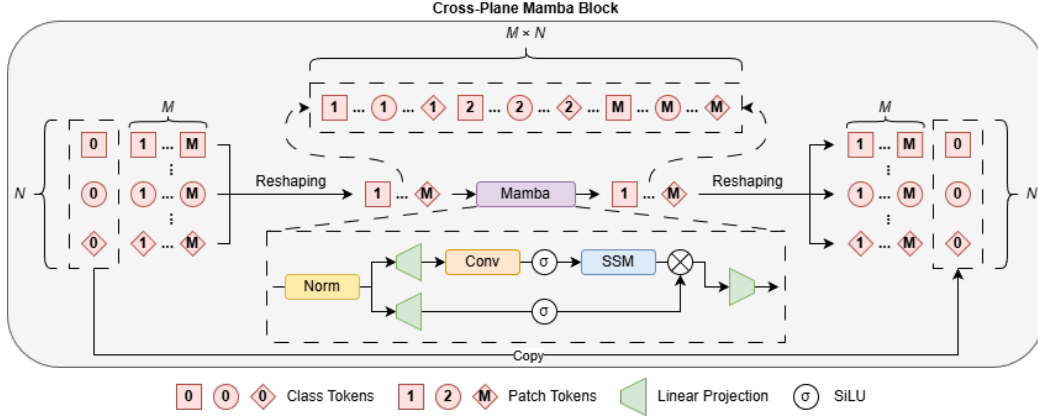
Figure 2: The CPM block comprises a vanilla Mamba block and two reshaping functions. The $N \times M$ patch tokens are separated from class tokens and reshaped into a single sequence of length $M \times N$, arranged such that neighboring tokens originate from different planes. Consequently, cross-plane modeling is achieved by learning interactions between the $M \times N$ tokens via SSM. The Mamba output is reshaped back into N sequences, each concatenated with its corresponding class token.

computation. This reshaping ensures that neighboring patch tokens originate from different planes, thereby prioritizing their interactions during cross-plane modeling:

$$\mathbf{x}^{'} = \text{Mamba}(\mathbf{x}) \tag{9}$$

The sequence $\mathbf{x}^{'}$ exiting the Mamba block is reshaped back into N parallel sequences, each consisting of M patch tokens and concatenated with its corresponding class token to form the input for the subsequent Transformer block.

The linear-time computation of Mamba ensures efficient learning of interactions between $M \times N$ tokens. Replacing SSM with SA for cross-plane modeling would significantly increase computational complexity, as detailed in Section 2.3.

Table 1: Batch size, sequence length, and computational complexity of in-plane and cross-plane modeling with SA or SSM. TranSamba settings are marked in gray.

	In-Plane	Cross-Plane
Sequence Length	M	MN
Batch Size	B	$\frac{B}{N}$
Time Complexity		
SA	$4MD^2 + 2M^2D$	$4MND^2 + 2M^2N^2D$
SSM	$128MD$	$128MND$
Space Complexity		
SA	$B(4MD^2 + 2M^2D)$	$\frac{B}{N}(4MND^2 + 2M^2N^2D)$
SSM	$B(128MD)$	$\frac{B}{N}(128MND)$

5

2.3. Computational Complexity Analysis

Given a sequence $\mathbf{T} \in \mathbb{R}^{1 \times L \times D}$, the time complexity of global modeling with SA is $4LD^2 + 2L^2D$; with SSM, when the expanded state dimension and the fixed parameter are set to $2D$ and 16 by default, the time complexity of global modeling is $128LD$ (Zhu et al., 2024).

Time Complexity. For a volume comprising N planes, the effective sequence lengths of in-plane and cross-plane modeling are M and MN (Table 1). With the series arrangement of the CPM block and the Transformer block in a TranSamba layer, the time complexity

$$128MND + 4MD^2 + 2M^2D \tag{10}$$

scales linearly with N. Replacing SSM with SA for cross-plane modeling would increase the time complexity, which scales quadratically with N:

$$4MND^2 + 2M^2N^2D + 4MD^2 + 2M^2D \tag{11}$$

Space Complexity. For a batch of B planes composed of $\frac{B}{N}$ volumes with N planes per volume, the effective batch sizes of in-plane and cross-plane modeling are B and $\frac{B}{N}$ (Table 1). The space complexity of TranSamba

$$B(4MD^2 + 2M^2D) + \frac{B}{N}(128MND) \tag{12}$$

is independent of N. Replacing SSM with SA for cross-plane modeling would cause the space complexity to increase with N:

$$B(4MD^2 + 2M^2D) + \frac{B}{N}(4MND^2 + 2M^2N^2D) \tag{13}$$

3. Experiments

3.1. Experimental Settings

Detailed experimental settings are presented in Appendix B.

Datasets. TranSamba is evaluated on three datasets: brain tumor segmentation (BraTS) (Menze et al., 2014; Bakas et al., 2017, 2018), kidney tumor segmentation (KiTS) (Heller et al., 2021), and left atrium segmentation challenge (LASC) (Xiong et al., 2021), comprising 484, 489, and 154 volumetric images, respectively. For each dataset, 100 volumetric images are held out for testing; the remaining images utilize an 80/20 training/validation split.

Evaluation. The 3D mean Dice similarity coefficient (DSC), 3D mean 95th percentile Hausdorff distance (HD), and 2D mean intersection-over-union (IoU) are used for evaluation. Ablation study results are reported on the validation set, while comparison with SOTA is reported on the test set.

Implementation. TranSamba is built on the DeiT-S (Touvron et al., 2021) pre-trained on ImageNet (Deng et al., 2009); the CPM blocks employ randomly initialized vanilla Mamba (Gu and Dao, 2024) blocks. Only slice-level labels are used for model training. For inference, the attention maps are generated in a single-stage manner.
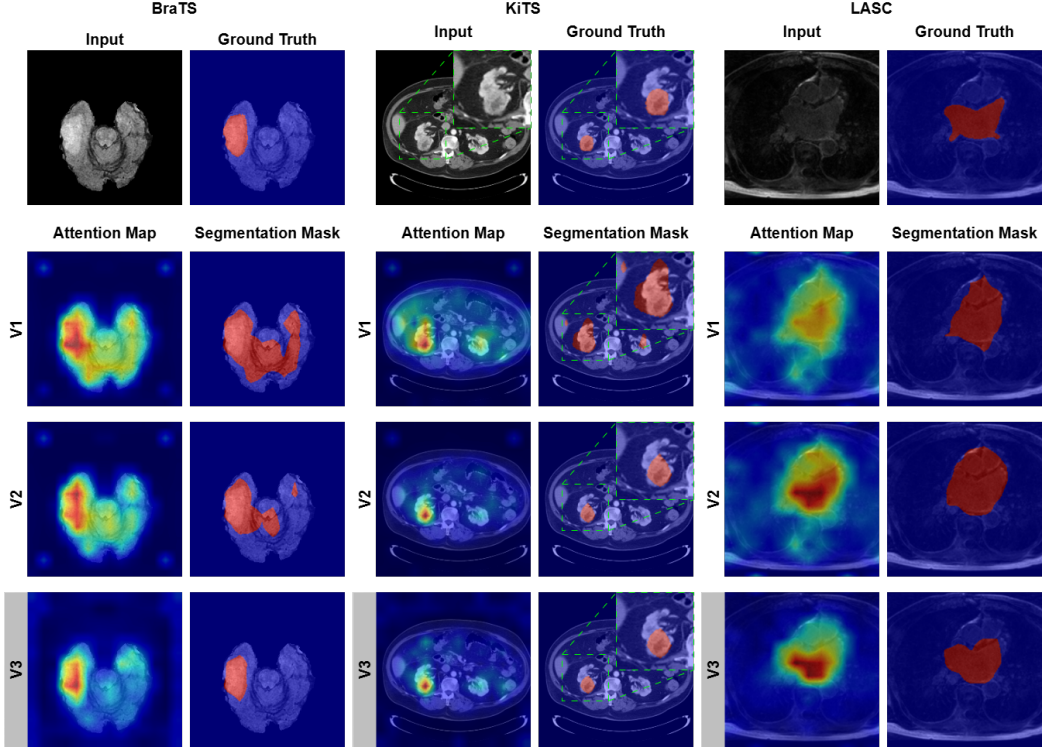
Figure 3: Qualitative results of ablation study on the effect of cross-plane modeling.

3.2. Ablation Studies

Ablation studies are conducted on the **1.** effect of cross-plane modeling, **2.** number of planes, **3.** hybrid layer design (Appendix C.1), and **4.** in-plane modeling methods (Appendix C.2).

Effect of Cross-Plane Modeling. We compare TranSamba with variants without Mamba blocks for cross-plane modeling, presenting quantitative results in Table 2. V3, representing TranSamba, achieves the highest performance across all datasets and evaluation metrics. V3 outperforms V1, a ViT-only baseline, by margins of 13.9%, 3.5 cm, and 9.7% in DSC, HD, and IoU averaged across datasets, respectively. To demonstrate that this

Table 2: Ablation study on the effect of cross-plane modeling. Variant performance is reported in DSC (%, ↑), HD (cm, ↓), and IoU (%, ↑). The highest performance is marked in **bold**, and the TranSamba variant is marked in gray.

Variant	Mamba	Memory (GB)	#params. (M)	BraTS			KiTS			LASC		
				DSC	HD	IoU	DSC	HD	IoU	DSC	HD	IoU
V1	-	13.0	21.7	40.7	5.6	28.8	15.8	17.4	20.1	33.5	10.4	20.4
V2	in-plane	17.6	33.3	41.9	5.1	28.2	19.5	16.6	24.7	38.1	10.7	24.3
V3	cross-plane	17.3	33.3	**60.7**	**4.6**	**44.2**	**25.1**	**15.2**	**24.9**	**45.9**	**3.2**	**29.3**

improvement does not result simply from scaling computational and model complexity, we introduce V2, comprising alternating Mamba and Transformer blocks similar to TranSamba. The Mamba blocks of V2 are equivalent to the forward branch in Vision Mamba (ViM) (Zhu et al., 2024), performing in-plane modeling alongside the Transformer blocks. Compared to V1, performance improves by incorporating additional Mamba blocks for in-plane modeling, particularly on the KiTS and LASC datasets. However, compared to V3, which shares the same model complexity but is marginally more efficient, the improvement attained by V2 is significantly lower. This suggests that the performance improvement of V3 primarily stems from information exchange between tokens from multiple neighboring planes rather than merely adding Mamba blocks. Qualitative comparisons of V1, V2, and V3 are presented in Figure 3. In the BraTS case, the tumor exhibits moderate contrast with normal brain tissue; both V1 and V2 attend to the entire brain area. With cross-plane modeling, V3 significantly reduces attention to normal brain tissue, thereby decreasing the false positive area. Similarly, in the LASC case, the contrast between the LA and other cardiac sub-structures is moderate. V1 localizes a large area covering the whole heart with moderate attention, while V2 attends more strongly to the LA but also to other substructures. In comparison, V3 focuses on the LA with significantly lower attention to the aorta and right ventricle, successfully avoiding most of them in the segmentation mask. In the KiTS case, the tumor attached to the right kidney is small. V1 attends to large background areas, including a section of the left kidney. Both V2 and V3 pinpoint the small tumor, with V3 yielding a marginally smaller false positive area. The superior quantitative and qualitative performance of V3 validates the effectiveness of cross-plane modeling.

Number of Planes. We evaluate the model performance when varying the number of planes per input volume, N, presenting quantitative results in Table 3. Regardless of N, incorporating cross-plane modeling consistently improves performance compared to in-plane-only baselines (Table 2). This demonstrates the robustness of cross-plane modeling to varying plane counts. By default, a TranSamba input volume comprises 16 planes. This default setting outperforms configurations with multiples of 16 planes across all datasets and metrics, excluding IoU on KiTS. This indicates that information exchange across neighboring planes is more beneficial than global information exchange.

Table 3: Ablation study on the number of planes. Performance is reported in DSC (%, ↑), HD (cm, ↓), and IoU (%, ↑). The highest performance is marked in **bold**, and the default setting for TranSamba is marked in gray. NA: Number of planes exceeds the minimal slice count of the dataset; implementation not applicable.

#planes	BraTS			KiTS			LASC		
	DSC	HD	IoU	DSC	HD	IoU	DSC	HD	IoU
128	57.6	5.0	43.6	NA	NA	NA	NA	NA	NA
64	55.1	5.1	41.8	20.0	17.4	24.8	NA	NA	NA
32	53.7	5.3	40.8	22.3	16.1	**27.7**	40.8	3.7	24.9
16	**60.7**	**4.6**	**44.2**	**25.1**	**15.2**	24.9	**45.9**	**3.2**	**29.3**

3.3. Comparison with State-of-the-Art

We compare TranSamba with SOTA methods for object localization and medical segmentation using slice-level labels. The selected methods are commonly used for comparison in related work and feature open implementation (Gildenblat and contributors, 2021). Moreover, these methods are not designed to model characteristics unique to a single type of anatomical object, ensuring generalizability across different pathologies.

The quantitative comparison with SOTA methods is presented in Table 4. TranSamba achieves the highest performance across all datasets and evaluation metrics, except for HD on the BraTS and KiTS datasets, where it ranks second. TranSamba outperforms the second-highest performer for each dataset by margins of 5.3% and 2.6% in DSC and IoU, averaged across datasets, respectively. Overall, TranSamba achieves the highest mean rank of 1.2 across datasets and metrics, demonstrating its superiority in weakly supervised volumetric medical segmentation.

TranSamba also demonstrates robustness in generalizing across datasets, being the only method consistently ranking in the top 50% across all datasets and metrics. All other methods fail to do so on at least one dataset. Grad-CAM (Selvaraju et al., 2017), FullGrad (Srinivas and Fleuret, 2019), and LayerCAM (Jiang et al., 2021) are general methods for object localization. Grad-CAM and LayerCAM perform well only on the BraTS and KiTS datasets, respectively, while FullGrad fails on the BraTS dataset despite performing well on the other two. AME-CAM (Chen et al., 2023b) and IAT (Lyu et al., 2024) were proposed for brain tumor segmentation. However, AME-CAM performs well only on the BraTS dataset, while IAT fails on the KiTS dataset. This robustness indicates that TranSamba successfully employs cross-plane modeling, a characteristic generalizable across pathologies and modalities, whereas other SOTA methods fail.

For each dataset, the qualitative results of the methods achieving the top-3 IoU are presented in Figure 4; a comprehensive qualitative comparison is provided in Appendix D. In the BraTS case, IAT and AME-CAM underestimate and overestimate the tumor area, respectively. In the KiTS case, FullGrad localizes a large area encompassing the tumor.

Table 4: Quantitative comparison with SOTA. Performance is reported in DSC (%, ↑), HD (cm, ↓), and IoU (%, ↑). For each metric, results are reported as value (rank). The mean rank of each method is averaged across all datasets and metrics. The highest, second, and third highest performance is marked in **bold**, *italic*, and <u>underlined</u>, respectively.

| Method	Backbone	Mean Rank	BraTS DSC	BraTS HD	BraTS IoU	KiTS DSC	KiTS HD	KiTS IoU	LASC DSC	LASC HD	LASC IoU
XGrad-CAM (Fu et al., 2020)	DeiT-S	9.3	11.2 (11)	6.5 (10)	6.5 (11)	3.2 (10)	20.3 (8)	3.2 (9)	4.1 (9)	11.8 (8)	2.3 (8)
Score-CAM (Wang et al., 2020)	DeiT-S	8.7	32.4 (8)	5.8 (9)	21.9 (7)	0.8 (11)	21.8 (10)	0.8 (11)	4.3 (8)	9.6 (5)	2.0 (9)
Eigen-CAM (Muhammad and Yeasin, 2020)	DeiT-S	7.1	25.3 (10)	6.5 (10)	16.1 (10)	5.8 (7)	19.9 (6)	6.2 (7)	6.7 (5)	9.1 (4)	3.4 (5)
Grad-CAM++ (Chattopadhay et al., 2018)	DeiT-S	7.1	45.9 (6)	5.4 (8)	28.3 (6)	3.8 (8)	21.9 (11)	5.6 (8)	8.8 (4)	12.0 (9)	5.9 (4)
AME-CAM (Chen et al., 2023b)	ResNet-18	7.0	*57.0 (2)*	4.3 (4)	*42.2 (2)*	3.5 (9)	19.3 (5)	2.3 (10)	1.1 (10)	14.0 (11)	0.6 (10)
Ablation-CAM (Ramaswamy et al., 2020)	DeiT-S	6.0	52.5 (5)	4.4 (5)	34.8 (5)	7.4 (6)	20.2 (7)	9.4 (6)	5.6 (7)	10.5 (6)	2.6 (7)
LayerCAM (Jiang et al., 2021)	DeiT-S	5.6	29.4 (9)	**2.5 (1)**	17.8 (9)	*21.4 (2)*	**13.5 (1)**	<u>17.3 (3)</u>	0.3 (11)	<u>6.0 (3)</u>	0.2 (11)
Grad-CAM (Selvaraju et al., 2017)	DeiT-S	5.0	<u>55.2 (3)</u>	<u>4.1 (3)</u>	36.6 (4)	10.7 (4)	21.1 (9)	<u>15.0 (4)</u>	6.0 (6)	<u>10.5 (6)</u>	2.8 (6)
IAT (Lyu et al., 2024)	DeiT-S	4.7	52.8 (4)	4.5 (6)	<u>37.1 (3)</u>	10.7 (4)	16.6 (4)	11.4 (5)	<u>36.1 (3)</u>	12.5 (10)	21.9 (3)
FullGrad (Srinivas and Fleuret, 2019)	DeiT-S	4.0	33.2 (7)	5.1 (7)	19.8 (8)	20.5 (3)	16.3 (3)	*25.1 (2)*	*37.7 (2)*	*4.3 (2)*	*23.1 (2)*
TranSamba (ours)	TranSamba	1.2	**64.1 (1)**	*3.9 (2)*	**46.8 (1)**	**26.8 (1)**	*14.4 (2)*	**25.8 (1)**	**41.0 (1)**	**4.1 (1)**	**25.6 (1)**
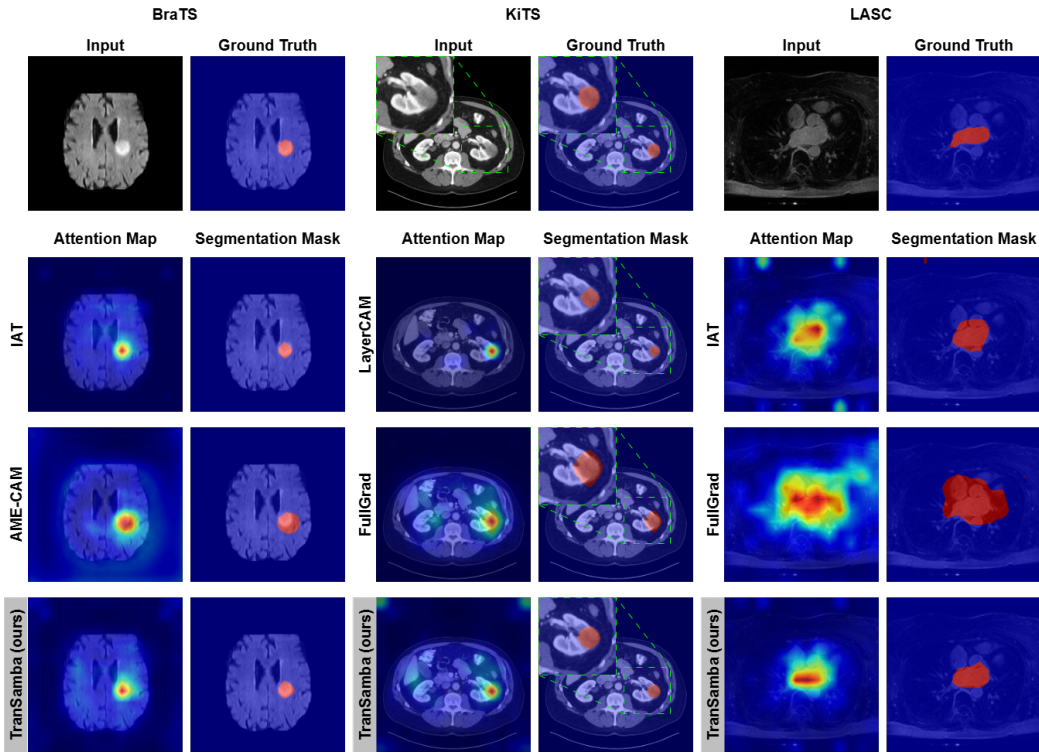
Figure 4: Qualitative comparison of methods achieving the top-3 IoU for each dataset.

Both LayerCAM and TranSamba pinpoint the small tumor, indicating their superiority in localizing small objects. In the LASC case, IAT attends to significant areas of the aorta and right ventricle, while FullGrad also includes a large area of the background in its segmentation. TranSamba attends to smaller areas of other substructures compared to IAT, resulting in a segmentation more conformal to the LA.

4. Conclusion

In this work, we introduce TranSamba, a hybrid architecture designed to bridge the gap between 2D WSSS and the volumetric nature of medical imaging. By integrating CPM blocks within a ViT encoder, TranSamba achieves efficient cross-plane modeling, thereby improving the quality of in-plane C2P attention for object localization. Theoretical analysis and empirical results confirm that our method maintains linear time complexity and constant memory usage with respect to the input volume depth. Extensive evaluation demonstrates that TranSamba significantly improves object localization accuracy compared to SOTA approaches, highlighting the generalizability of cross-plane modeling across diverse segmentation tasks. These findings underscore the efficacy of combining cross-plane SSM with in-plane SA for weakly supervised volumetric medical segmentation. Future work could extend TranSamba to multi-class scenarios as well as object localization in videos.

Acknowledgments

References

Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022.

Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4(1):1–13, 2017.

Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint arXiv:1811.02629*, 2018.

Ao Chang, Jiajun Zeng, Ruobing Huang, and Dong Ni. Em-net: Efficient channel and frequency learning with mamba for 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 266–275. Springer, 2024.

Ruiheng Chang, Dong Wang, Haiyan Guo, Jia Ding, and Liwei Wang. Weakly-supervised ultrasound video segmentation with minimal annotations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 648–658. Springer, 2021.

Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE, 2018.

Huazhen Chen, Jianpeng An, Bochang Jiang, Lili Xia, Yunhao Bai, and Zhongke Gao. Wsmtst: Weakly supervised multi-label brain tumor segmentation with transformers. *IEEE Journal of Biomedical and Health Informatics*, 27(12):5914–5925, 2023a.

Yu-Jen Chen, Xinrong Hu, Yiyu Shi, and Tsung-Yi Ho. Ame-cam: Attentive multiple-exit cam for weakly supervised segmentation on mri brain tumor. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 173–182. Springer, 2023b.

Zhang Chen, Zhiqiang Tian, Jihua Zhu, Ce Li, and Shaoyi Du. C-cam: Causal cam for weakly supervised semantic segmentation on medical image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11676–11685, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Reuben Dorent, Samuel Joutard, Jonathan Shapey, Aaron Kujawa, Marc Modat, Sébastien Ourselin, and Tom Vercauteren. Inter extreme points geodesics for end-to-end weakly supervised image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 615–624. Springer, 2021.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Hao Du, Qihua Dong, Yan Xu, and Jing Liao. Weakly-supervised 3d medical image segmentation using geometric prior and contrastive similarity. *IEEE Transactions on Medical Imaging*, 42(10):2936–2947, 2023.

Jiansong Fan, Tianxu Lv, Yicheng Di, Lihua Li, and Xiang Pan. Pathmamba: Weakly supervised state space model for multi-class segmentation of pathology images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 500–509. Springer, 2024.

Jia Fu, Tao Lu, Shaoting Zhang, and Guotai Wang. Um-cam: Uncertainty-weighted multi-resolution class activation maps for weakly-supervised fetal brain segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 315–324. Springer, 2023.

Ruigang Fu, Qingyong Hu, Xiaohu Dong, Yulan Guo, Yinghui Gao, and Biao Li. Axiom-based grad-cam: Towards accurate visualization and explanation of cnns. *arXiv preprint arXiv:2008.02312*, 2020.

Feng Gao, Minhao Hu, Min-Er Zhong, Shixiang Feng, Xuwei Tian, Xiaochun Meng, Zeping Huang, Minyi Lv, Tao Song, Xiaofan Zhang, et al. Segmentation only uses sparse annotations: Unified weakly and semi-supervised learning in medical images. *Medical Image Analysis*, 80:102515, 2022.

Jacob Gildenblat and contributors. Pytorch library for cam methods. https://github.com/jacobgil/pytorch-grad-cam, 2021.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First conference on language modeling*, 2024.

Ali Hatamizadeh and Jan Kautz. Mambavision: A hybrid mamba-transformer vision backbone. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25261–25270, 2025.

Nicholas Heller, Fabian Isensee, Klaus H Maier-Hein, Xiaoshuai Hou, Chunmei Xie, Fengyi Li, Yang Nan, Guangrui Mu, Zhiyong Lin, Miofei Han, et al. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. *Medical image analysis*, 67:101821, 2021.

Xinrong Hu, Yu-Jen Chen, Tsung-Yi Ho, and Yiyu Shi. Conditional diffusion models for weakly supervised medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 756–765. Springer, 2023.

Chao Huang, Yushu Shi, Waikeung Wong, Chengliang Liu, Wei Wang, Zhihua Wang, and Jie Wen. Multi-view evidential learning-based medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 17386–17394, 2025a.

Jiahao Huang, Liutao Yang, Fanwen Wang, Yinzhe Wu, Yang Nan, Weiwen Wu, Chengyan Wang, Kuangyu Shi, Angelica I Aviles-Rivero, Carola-Bibiane Schoenlieb, et al. Enhancing global sensitivity and uncertainty quantification in medical image reconstruction with monte carlo arbitrary-masked mamba. *Medical Image Analysis*, 99:103334, 2025b.

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.

Zexin Ji, Beiji Zou, Xiaoyan Kui, Pierre Vera, and Su Ruan. Deform-mamba network for mri super-resolution. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 242–252. Springer, 2024.

Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. Layercam: Exploring hierarchical class activation maps for localization. *IEEE transactions on image processing*, 30:5875–5888, 2021.

Arunkumar Kannan, Martin A Lindquist, and Brian Caffo. Brainmt: A hybrid mamba-transformer architecture for modeling long-range dependencies in functional mri data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 150–160. Springer, 2025.

Abbas Khan, Muhammad Asad, Martin Benning, Caroline Roney, and Gregory Slabaugh. Cams: Convolution and attention-free mamba-based cardiac image segmentation. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1893–1903. IEEE, 2025.

Alexander Kolesnikov and Christoph H Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *European conference on computer vision*, pages 695–711. Springer, 2016.

Yilmaz Korkmaz and Vishal M Patel. Mambarecon: Mri reconstruction with structured state space models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4142–4152. IEEE, 2025.

Zhuo Kuang, Zengqiang Yan, Huiyu Zhou, and Li Yu. Cluster-re-supervision: Bridging the gap between image-level and pixel-wise labels for weakly supervised medical image segmentation. *IEEE Journal of Biomedical and Health Informatics*, 27(10):4890–4901, 2023.

Issam Laradji, Pau Rodriguez, Oscar Manas, Keegan Lensink, Marco Law, Lironne Kurzman, William Parker, David Vazquez, and Derek Nowrouzezahrai. A weakly supervised consistency-based learning method for covid-19 segmentation in ct images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2453–2462, 2021.

Wenhui Lei, Qi Su, Tianyu Jiang, Ran Gu, Na Wang, Xinglong Liu, Guotai Wang, Xiaofan Zhang, and Shaoting Zhang. One-shot weakly-supervised segmentation in 3d medical images. *IEEE Transactions on Medical Imaging*, 43(1):175–189, 2023.

Marvin Lerousseau, Maria Vakalopoulou, Marion Classe, Julien Adam, Enzo Battistella, Alexandre Carré, Théo Estienne, Théophraste Henry, Eric Deutsch, and Nikos Paragios. Weakly supervised multiple instance learning histopathological tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 470–479. Springer, 2020.

Kailu Li, Ziniu Qian, Yingnan Han, Eric I-Chao Chang, Bingzheng Wei, Maode Lai, Jing Liao, Yubo Fan, and Yan Xu. Weakly supervised histopathology image segmentation with self-attention. *Medical Image Analysis*, 86:102791, 2023.

Yongshuai Li, Yuan Liu, Lijie Huang, Zhili Wang, and Jianwen Luo. Deep weakly-supervised breast tumor segmentation in ultrasound images with explicit anatomical constraints. *Medical image analysis*, 76:102315, 2022.

Zhe Li and Yong Xia. Deep reinforcement learning for weakly-supervised lymph node segmentation in ct images. *IEEE Journal of Biomedical and Health Informatics*, 25(3): 774–783, 2020.

Pengchen Liang, Leijun Shi, Bin Pu, Renkai Wu, Jianguo Chen, Lixin Zhou, Lite Xu, Zhuangzhuang Chen, Qing Chang, and Yiwei Li. Mambasam: A visual mamba-adapted sam framework for medical image segmentation. *IEEE Journal of Biomedical and Health Informatics*, 2025.

Opher Lieber, Barak Lenz, Hofit Bata, Gal Cohen, Jhonathan Osin, Itay Dalmedigos, Erez Safahi, Shaked Meirom, Yonatan Belinkov, Shai Shalev-Shwartz, et al. Jamba: A hybrid transformer-mamba language model. *arXiv preprint arXiv:2403.19887*, 2024.

Anglin Liu, Dengqiang Jia, Kaicong Sun, Runqi Meng, Meixin Zhao, Yongluo Jiang, Zhijian Dong, Yaozong Gao, and Dinggang Shen. Lm-unet: Whole-body pet-ct lesion segmentation with dual-modality-based annotations driven by latent mamba u-net. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 405–414. Springer, 2024a.

Jiarun Liu, Hao Yang, Hong-Yu Zhou, Yan Xi, Lequan Yu, Cheng Li, Yong Liang, Guang-ming 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, 2024b.

Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. Vmamba: Visual state space model. *Advances in neural information processing systems*, 37:103031–103063, 2024c.

Yunze Liu and Li Yi. Map: Unleashing hybrid mamba-transformer vision backbone's potential with masked autoregressive pretraining. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9676–9685, 2025.

Yiheng Lyu, Lian Xu, Mohammed Bennamoun, Farid Boussaid, and Girish Dwivedi. Importance-aware transformer: Addressing intra-class heterogeneity in weakly supervised brain tumor segmentation. In *2024 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 300–307. IEEE, 2024.

Jun Ma, Feifei Li, and Bo Wang. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722*, 2024.

Xiao Ma, Zexuan Ji, Sijie Niu, Theodore Leng, Daniel L Rubin, and Qiang Chen. Mscam: Multi-scale class activation maps for weakly-supervised segmentation of geographic atrophy lesions in sd-oct images. *IEEE journal of biomedical and health informatics*, 24 (12):3443–3455, 2020.

Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016.

Mohammed Bany Muhammad and Mohammed Yeasin. Eigen-cam: Class activation map using principal components. In *2020 international joint conference on neural networks (IJCNN)*, pages 1–7. IEEE, 2020.

Harish Guruprasad Ramaswamy et al. Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. In *proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 983–991, 2020.

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.

Marius Schmidt-Mengin, Alexis Benichoux, Shibeshih Belachew, Nikos Komodakis, and Nikos Paragios. Tonno: Tomographic reconstruction of a neural network's output for

weakly supervised segmentation of 3d medical images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11428–11438, 2024.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

Amber L Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram Van Ginneken, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*, 2019.

Suraj Srinivas and François Fleuret. Full-gradient representation for neural network visualization. *Advances in neural information processing systems*, 32, 2019.

Fenghe Tang, Bingkun Nian, Yingtai Li, Zihang Jiang, Jie Yang, Wei Liu, and S Kevin Zhou. Mambamim: Pre-training mamba with state space token interpolation and its application to medical image segmentation. *Medical Image Analysis*, page 103606, 2025.

Wen Tang, Han Kang, Ying Cao, Pengxin Yu, Hu Han, Rongguo Zhang, and Kuan Chen. M-seam-nam: multi-instance self-supervised equivalent attention mechanism with neighborhood affinity module for double weakly supervised segmentation of covid-19. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 262–272. Springer, 2021a.

Youbao Tang, Jinzheng Cai, Ke Yan, Lingyun Huang, Guotong Xie, Jing Xiao, Jingjing Lu, Gigin Lin, and Le Lu. Weakly-supervised universal lesion segmentation with regional level set loss. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 515–525. Springer, 2021b.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.

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.

Gui Wang, Yuexiang Li, Wenting Chen, Meidan Ding, Wooi Ping Cheah, Rong Qu, Jianfeng Ren, and Linlin Shen. S^3-mamba: Small-size-sensitive mamba for lesion segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 7655–7664, 2025a.

Han Wang, Fasheng Yi, Jingling Wang, Zhang Yi, and Haixian Zhang. Recistsup: Weakly-supervised lesion volume segmentation using recist measurement. *IEEE Transactions on Medical Imaging*, 41(7):1849–1861, 2022.

Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 24–25, 2020.

Haoran Wang, Lian Huai, Wenbin Li, Lei Qi, Xingqun Jiang, and Yinghuan Shi. Weakmed-sam: Weakly-supervised medical image segmentation via sam with sub-class exploration and prompt affinity mining. *IEEE Transactions on Medical Imaging*, 2025b.

Hualiang Wang, Yiqun Lin, Xinpeng Ding, and Xiaomeng Li. Tri-plane mamba: Efficiently adapting segment anything model for 3d medical images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 636–646. Springer, 2024a.

Jinhong Wang, Jintai Chen, Danny Chen, and Jian Wu. Lkm-unet: Large kernel vision mamba unet for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 360–370. Springer, 2024b.

Juan Wang and Bin Xia. Bounding box tightness prior for weakly supervised image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 526–536. Springer, 2021.

Udaranga Wickramasinghe, Patrick Jensen, Mian Shah, Jiancheng Yang, and Pascal Fua. Weakly supervised volumetric image segmentation with deformed templates. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 422–432. Springer, 2022.

Zhaohu Xing, Tian Ye, Yijun Yang, Guang Liu, and Lei Zhu. Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 578–588. Springer, 2024.

Zhaohan Xiong, Qing Xia, Zhiqiang Hu, Ning Huang, Cheng Bian, Yefeng Zheng, Sulaiman Vesal, Nishant Ravikumar, Andreas Maier, Xin Yang, et al. A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging. *Medical image analysis*, 67:101832, 2021.

Lian Xu, Wanli Ouyang, Mohammed Bennamoun, Farid Boussaid, and Dan Xu. Multi-class token transformer for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4310–4319, 2022.

Lian Xu, Mohammed Bennamoun, Farid Boussaid, Hamid Laga, Wanli Ouyang, and Dan Xu. Mctformer+: Multi-class token transformer for weakly supervised semantic segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 46(12):8380–8395, 2024.

Jiaqi Yang, Nitish Mehta, Gozde Demirci, Xiaoling Hu, Meera S Ramakrishnan, Mina Naguib, Chao Chen, and Chia-Ling Tsai. Anomaly-guided weakly supervised lesion segmentation on retinal oct images. *Medical image analysis*, 94:103139, 2024a.

Zefan Yang, Jiajin Zhang, Ge Wang, Mannudeep K Kalra, and Pingkun Yan. Cardiovascular disease detection from multi-view chest x-rays with bi-mamba. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 134–144. Springer, 2024b.

Dong Zhang, Bo Chen, Jaron Chong, and Shuo Li. Weakly-supervised teacher-student network for liver tumor segmentation from non-enhanced images. *Medical Image Analysis*, 70:102005, 2021.

Hao Zheng, Zhiguo Zhuang, Yulei Qin, Yun Gu, Jie Yang, and Guang-Zhong Yang. Weakly supervised deep learning for breast cancer segmentation with coarse annotations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 450–459. Springer, 2020.

Yuming Zhong and Yi Wang. Simple: Similarity-aware propagation learning for weakly-supervised breast cancer segmentation in dce-mri. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 567–577. Springer, 2023.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

Lingyu Zhou, Zhang Yi, Kai Zhou, and Xiuyuan Xu. Efficient and gender-adaptive graph vision mamba for pediatric bone age assessment. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 230–239. Springer, 2024.

Meng Zhou, Zhe Xu, Kang Zhou, and Raymond Kai-yu Tong. Weakly supervised medical image segmentation via superpixel-guided scribble walking and class-wise contrastive regularization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 137–147. Springer, 2023.

Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024.

Meilu Zhu, Zhen Chen, and Yixuan Yuan. Feddm: Federated weakly supervised segmentation via annotation calibration and gradient de-conflicting. *IEEE Transactions on Medical Imaging*, 42(6):1632–1643, 2023.

Appendix A. Related Work

A.1. Weakly Supervised Medical Segmentation

Various forms of weak supervision have been explored for medical segmentation. Common types include **1.** point labels, marking single pixels per region (Laradji et al., 2021) or boundary points (Dorent et al., 2021; Wickramasinghe et al., 2022; Zhong and Wang, 2023); **2.** scribble labels, providing coarse contours (Zheng et al., 2020) or sparse labels (Gao et al., 2022; Lei et al., 2023; Zhou et al., 2023); **3.** bounding-box labels, encompassing targets

either loosely (Zhu et al., 2023) or tightly (Wang and Xia, 2021; Zhang et al., 2021; Du et al., 2023). Alternatively, clinical workflow metrics, such as diameter measurements from tumor response criteria, can be repurposed for WSSS (Li and Xia, 2020; Tang et al., 2021b; Wang et al., 2022). However, such metrics are typically restricted to malignant tumors, limiting their applicability.

Compared to these forms, slice-level labels are highly cost-effective. However, they present significant challenges for model training due to the sparsity of localization information. Previous research has focused on distinct strategies to address the specific challenges associated with slice-level supervision.

Multi-Scale Feature Aggregation. Feature maps in convolutional neural networks undergo progressive down-sampling, producing low-resolution CAMs that often fail to capture fine structural details. This issue is particularly critical in tasks requiring high precision, such as retinal optical coherence tomography (OCT) or brain MRI. To mitigate this, several studies extract CAMs at multiple resolutions (Ma et al., 2020; Chen et al., 2023b; Fu et al., 2023). Rather than simple summation, these methods employ sophisticated aggregation strategies to effectively fuse information across scales, such as an attentional fully connected layer (Ma et al., 2020), an unsupervised contrastive attention network (Chen et al., 2023b), or uncertainty weighting (Fu et al., 2023).

Handling Challenging Object Characteristics. Anatomical objects often exhibit challenging appearances. For instance, low contrast resolution leads to poorly defined boundaries (Chang et al., 2021; Chen et al., 2022). Additionally, objects can be extremely small relative to the image size; examples of such objects include breast tumors in ultrasound (Chang et al., 2021) or lesions in retinal OCT (Yang et al., 2024a). To address this, Chang et al. (Chang et al., 2021) enforce consistency between co-attentive features from ultrasound video frames and their corresponding CAMs. C-CAM (Chen et al., 2022) introduces causal intervention to focus attention on foreground regions with ambiguous boundaries. Yang et al. (Yang et al., 2024a) combines generative anomaly detection with self-attention to capture global context for small-lesion segmentation.

Self-Supervised Regularization. Self-supervised learning is frequently integrated with WSSS to bridge the gap between slice-level supervision and voxel-level segmentation. Inspired by its success in natural images, the SEAM framework has been adapted for medical segmentation, often with additional regularization such as neighborhood affinity constraints (Tang et al., 2021a) or weighted in-plane and cross-plane importance (Lyu et al., 2024). Kuang et al. (Kuang et al., 2023) leverage inter-channel reconstruction via random masking for self-supervision, complemented by unsupervised clustering to generate pixel-level constraints from slice-level labels.

Incorporating Domain-Specific Priors. Another line of research incorporates domain-specific constraints. For instance, C-CAM (Chen et al., 2022) applies anatomical priors to manage co-occurring foreground objects of different classes. Li et al. (Li et al., 2022) constrain breast tumor activations to specific tissue layers and introduce a deep level-set loss to encourage smooth boundaries and homogeneous intensities. For brain tumor sub-regions, WS-MTST (Chen et al., 2023a) proposes aggregation and overlap losses to promote clustered activations and preserve sub-region relationships.

Other approaches explore different supervision levels or backbone architectures. ToNNO (Schmidt-Mengin et al., 2024) generates 3D localization maps from a 2D encoder via volu-

metric reconstruction using volume-level labels. However, volume-level supervision is generally inapplicable to normal structures where negative samples, for example, cardiac CT studies that entirely lack a cardiac chamber, do not naturally exist. Other notable efforts employ diffusion models (Hu et al., 2023) or foundation models (Wang et al., 2025b) for slice-level WSSS. While image-level supervision has shown utility in multiple instance learning for histopathology segmentation (Lerousseau et al., 2020; Li et al., 2023; Fan et al., 2024), the fundamental differences between histopathology and volumetric imaging place it beyond the scope of this work.

A.2. Mamba in Medical Imaging

Originally proposed for sequence modeling in domains such as audio, genomics, and natural language processing, Mamba (Gu and Dao, 2024) has been adapted to computer vision, pioneered by ViM (Zhu et al., 2024) and VMamba (Liu et al., 2024c). Following the plain design of ViT (Dosovitskiy, 2020), ViM replaces Transformer layers with bidirectional SSM layers that scan tokens in forward and backward directions to mitigate the unidirectional nature of SSM. Conversely, VMamba features a hierarchical design with decreasing resolution and increasing channel dimension, utilizing a quadridirectional cross-scan mechanism to integrate information from spatially neighboring patches. Recently, hybrid architectures have emerged to leverage both the linear-time complexity of SSM and the powerful attention mechanism. Jamba (Lieber et al., 2024) validates this synergy for language modeling, prompting similar advances in computer vision. MambaVision (Hatamizadeh and Kautz, 2025) introduces a hierarchical architecture leveraging convolutional blocks in early high-resolution stages and hybrid Mamba-Transformer blocks in deeper stages. These hybrid blocks feature a redesigned MambaVision mixer and SA for enhanced long-range modeling. Complementing architectural innovation, MAP (Liu and Yi, 2025) proposes a masked autoregressive pretraining strategy tailored for hybrid models, effectively unifying the next-token prediction of Mamba with the masked signal reconstruction of Transformer.

In medical imaging, Mamba has been applied to diverse tasks including image super-resolution (Ji et al., 2024) and reconstruction (Huang et al., 2025b; Korkmaz and Patel, 2025). Another important application of Mamba is computer-aided diagnosis (Yang et al., 2024b; Zhou et al., 2024; Kannan et al., 2025). For instance, BI-Mamba (Yang et al., 2024b) proposes a bidirectional SSM-based model for cardiovascular disease detection from chest X-ray. Notably, BrainMT (Kannan et al., 2025) proposes a hybrid Mamba-Transformer architecture for long-range modeling in functional MRI phenotyping.

Segmentation remains the most common application of Mamba in medical imaging, primarily in a fully supervised manner. Following the success of U-Net (Ronneberger et al., 2015), numerous Mamba-based medical segmentation methods (Ma et al., 2024; Xing et al., 2024; Liu et al., 2024b; Wang et al., 2024b; Liu et al., 2024a) adopt a U-shaped design. U-Mamba (Ma et al., 2024) integrates convolutional blocks with Mamba blocks; convolutional layers extract feature maps while Mamba models long-range dependency, addressing the limited local receptive field of convolutional kernels. SegMamba (Xing et al., 2024) combines 3D convolutional kernels with a novel tri-oriented Mamba block that captures both cross-plane and bidirectional in-plane interactions within 3D feature maps. Swin-UMamba (Liu et al., 2024b) and MambaMIM (Tang et al., 2025) further demonstrate the benefits

of pre-training on large-scale natural or medical datasets, achieving strong downstream segmentation performance.

Specialized methods have also been developed to address specific challenges. EM-Net (Chang et al., 2024) leverages Mamba to enhance feature learning in both spatial and frequency domains, capturing complementary representations. CAMS-Net (Khan et al., 2025) removes the convolutional operation in the vanilla Mamba block, reducing complexity without compromising segmentation accuracy. S^3-Mamba (Wang et al., 2025a) introduces a size-aware model design coupled with a tailored training strategy to address the localization of small lesions. Furthermore, Mamba-based blocks have been integrated with foundation models to improve adaptation to the medical domain (Wang et al., 2024a; Huang et al., 2025a; Liang et al., 2025). In weakly supervised settings, PathMamba (Fan et al., 2024) leverages Mamba-based blocks to capture patch-level dependency, which is used to construct a self-supervised contrastive loss enforcing consistency between patch-level representations and pixel-level feature maps.

Appendix B. Supplementary Experimental Settings

B.1. Datasets

TranSamba is evaluated on three public datasets covering distinct pathologies and imaging modalities to demonstrate the generalizability of cross-plane modeling. All volumetric images are processed from the axial view, with axial slices extracted as 2D planes. Following nnU-Net (Isensee et al., 2021) we pre-process the volumetric images, including cropping or padding to an in-plane spatial resolution of $(224, 224)$. Table 5 summarizes dataset characteristics after pre-processing. Slice-level labels are set to positive if the slice contains the segmentation target, regardless of size.

Brain Tumor Segmentation. We use the training set of the Medical Segmentation Decathlon (Simpson et al., 2019; Antonelli et al., 2022) BraTS dataset (Menze et al., 2014; Bakas et al., 2017, 2018), comprising 484 multiparametric MRI studies of patients with glioma. The T2 fluid-attenuated inversion recovery (FLAIR) sequence is utilized, with the segmentation target defined as the whole tumor, encompassing edema, enhancing, and non-enhancing regions. Intensity normalization is performed using z-scoring.

Kidney Tumor Segmentation. We use the training set of the 2023 KiTS (Heller et al., 2021) dataset, containing 489 contrast-enhanced CT studies of patients with suspected kidney tumors. The segmentation target is the kidney tumor. All volumetric images are **1.** clipped to an abdominal CT window of [-150, 250] Hounsfield units, **2.** resampled to an isotropic voxel size of 2 mm^3, and **3.** normalized using the global foreground mean and standard deviation. Compared to the abdominal cavity, the kidney tumor volume is

Table 5: Key characteristics of the BraTS, the KiTS, and the LASC datasets.

Dataset	BraTS	KiTS	LASC
Target	Whole Tumor	Tumor	LA Cavity
Modality	T2-FLAIR MRI	CT	LGE-MRI
#studies	484	489	154
min. #slices	155	71	44

typically small, causing severe slice-level imbalance. To mitigate this, slices located at least 3.2 cm away from both the kidney and the tumor are discarded, provided that the number of remaining slices in each study was no fewer than 71, the minimum slice count of the dataset.

Left Atrium Segmentation Challenge. We use both the training and testing sets of the 2018 LASC (Xiong et al., 2021) dataset, comprising 154 late gadolinium-enhanced (LGE) MRI studies of patients with atrial fibrillation. The segmentation target is the LA cavity. All volumetric images are resampled to an isotropic voxel size of 1.25 mm^3 and normalized using z-scoring.

These datasets validate TranSamba under diverse and challenging scenarios. BraTS presents complex, heterogeneously located gliomas. KiTS includes both small tumors embedded within a large abdominal cavity and extremely large tumors, posing scale imbalance and model bias challenges. LASC spans a wide range of image quality; additionally, only 43 volumetric studies are assigned to the training set, corresponding to fewer than 2,000 2D images.

By fusing sub-regions in BraTS and targeting only the tumor in KiTS, we formulate the task as binary WSSS. While multi-class WSSS is important, we argue that such tasks require methods orthogonal to cross-plane modeling, such as learning inter-class (Xu et al., 2022, 2024) or inter-channel (Kuang et al., 2023) correlation, addressing object co-occurrence (Chen et al., 2022), or using customized loss functions (Chen et al., 2023a).

B.2. Implementation

During training, we randomly sample volumes containing 16 contiguous planes from the input volumetric images. Each batch comprises 16 volumes, resulting in a batch size of 256 planes. All models are trained for 100 epochs without early stopping. The BCE loss is weighted to address the imbalance between positive and negative planes. We follow MCTformer+ (Xu et al., 2024) for data augmentation and other training hyperparameters. For inference, volumes comprising 16 continuous planes are sampled sequentially from the input volumetric images. Final segmentation masks are obtained by applying a binary threshold of 0.5 to attention maps generated directly on the validation and testing sets. No post-processing is applied, demonstrating the intrinsic quality of object localization.

Table 6: Ablation study on the hybrid layer design. Performance is reported in DSC (%, ↑), HD (cm, ↓), and IoU (%, ↑). For each metric, results are reported as value (rank). The mean rank of each method is averaged across all datasets and metrics. The highest performance is marked in **bold**, and the default design for TranSamba is marked in gray.

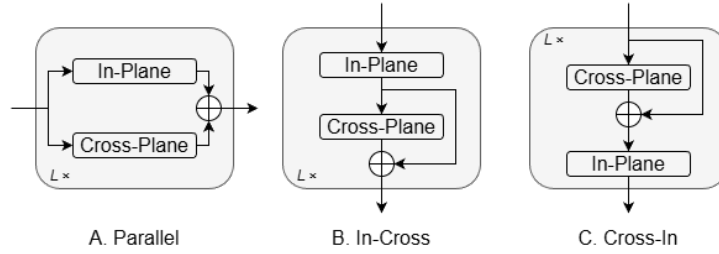| Design	Mean Rank	BraTS			KiTS			LASC		
		DSC	HD	IoU	DSC	HD	IoU	DSC	HD	IoU
Parallel	2.2	60.8 (2)	4.8 (3)	42.0 (3)	23.9 (3)	15.5 (3)	24.4 (3)	**47.6 (1)**	**3.1 (1)**	**30.9 (1)**
In-Cross	1.9	**64.4 (1)**	**3.8 (1)**	44.2 (2)	24.6 (2)	**15.4 (1)**	**25.2 (1)**	42.0 (3)	3.6 (3)	26.2 (3)
Cross-In	1.8	60.7 (3)	4.6 (2)	**44.2 (1)**	**25.1 (1)**	**15.2 (1)**	24.9 (2)	45.9 (2)	3.2 (2)	29.3 (2)

Figure 5: Different designs of the hybrid layer. Design A: Parallel, cross-plane block in parallel with the in-plane block. Design B: In-Cross, in-plane block in series before the cross-plane block. Design C: Cross-In, cross-plane block in series before the in-plane block. All designs use vanilla Transformer and CPM blocks for in-plane and cross-plane blocks, respectively. Residual connections are embedded within the in-plane Transformer blocks.

Training and inference are conducted on a single NVIDIA A100-SXM4 GPU with 40 GB of memory.

Appendix C. Supplementary Ablation Studies

C.1. Hybrid Layer Design

We compare model performance with three different hybrid layer designs, as shown in Figure 5. In Design A, each hybrid layer comprises a pair of parallel in-plane and cross-plane blocks. For Designs B and C, each layer consists of a pair of in-plane and cross-plane blocks in series, with a residual connection placed between the two blocks. The cross-plane block is placed after the in-plane block in Design B and before the in-plane block in Design C. For all designs, the cross-plane block is a CPM block, while the in-plane block is a vanilla Transformer block with an embedded residual connection.

The quantitative comparison of hybrid layer designs is presented in Table 6. Regardless of the layer design, incorporating cross-plane modeling consistently improves performance

Table 7: Ablation study on the in-plane modeling method. Variant performance is reported in DSC (%, ↑), HD (cm, ↓), and IoU (%, ↑). The highest performance is marked in **bold**, and the TranSamba variant is marked in gray. The memory usage of V5 is empirically estimated rather than experimentally measured.

Variant	Cross-Plane	In-Plane	Memory (GB)	#params. (M)	BraTS DSC	HD	IoU	KiTS DSC	HD	IoU	LASC DSC	HD	IoU
V1	-	SA	13.0	21.7	40.7	5.6	28.8	15.8	17.4	20.1	33.5	10.4	20.4
V2	-	Bi-SSM	9.2	23.6	32.6	5.7	18.3	5.8	15.9	5.3	0.3	12.8	0.1
V3	SSM	SA	17.3	33.3	**60.7**	**4.6**	**44.2**	**25.1**	**15.2**	**24.9**	**45.9**	**3.2**	**29.3**
V4	SSM	Bi-SSM	13.5	35.2	36.7	5.0	19.5	6.0	15.8	5.5	1.4	12.5	0.6
V5	SA	-	52.6	21.7	-	-	-	-	-	-	-	-	-

compared to in-plane-only baselines (Table 2). This demonstrates the robustness of cross-plane modeling to different layer designs. By default, TranSamba employs Design C, which marginally outperforms Design B, followed by Design A. Compared to its counterparts, Design C computes in-plane SA following cross-plane modeling via SSM. This design ensures that in each layer, C2P attention is extracted following information exchange across neighboring planes, thereby maximizing the quality improvement of the attention maps.

C.2. In-Plane Modeling Methods

We compare variants employing different methods for in-plane modeling, with and without SSM-based cross-plane modeling, presenting quantitative results in Table 7. V1 and V3 represent a ViT-only baseline and TranSamba, respectively. For V2 and V4, we replace the Transformer block in each hybrid layer with a ViM-like bidirectional SSM block. In the absence of pairwise SA, localization maps are generated from patch token-based CAMs (Xu et al., 2022).

Compared to V1 and V3, using bidirectional SSM blocks for in-plane modeling results in significantly inferior performance across all datasets and evaluation metrics. Notably, V2 and V4 fail on the more challenging KiTS and LASC datasets. The performance drop observed justifies the hybrid design of TranSamba. Nevertheless, incorporating cross-plane modeling consistently improves performance regardless of the in-plane modeling method.

For completeness, we also provide the memory usage for V5, which employs SA for cross-plane modeling. Memory usage is estimated from the space complexity of cross-plane modeling with SA detailed in Table 1, assuming 16 planes per input volume. The estimated memory exceeds available computational resources and would continue to grow as the input volume depth increases.

Appendix D. Qualitative Comparison with State-of-the-Art

The comprehensive qualitative comparison with SOTA methods is presented in Figure 6 and Figure 7. In the BraTS case, most methods successfully localize the tumor; exceptions include XGrad-CAM (Fu et al., 2020), which primarily activates the background, and Eigen-CAM (Muhammad and Yeasin, 2020), which activates the entire brain area. Similarly, in the KiTS case, Eigen-CAM activates the entire body area. Only four methods, Grad-CAM++ (Chattopadhay et al., 2018), LayerCAM (Jiang et al., 2021), FullGrad (Srinivas and Fleuret, 2019), and TranSamba, successfully localize the small tumor attached to the left kidney. The LASC case also presents a significant challenge, with only Grad-CAM++, IAT (Lyu et al., 2024), FullGrad, and TranSamba successfully localizing the LA.
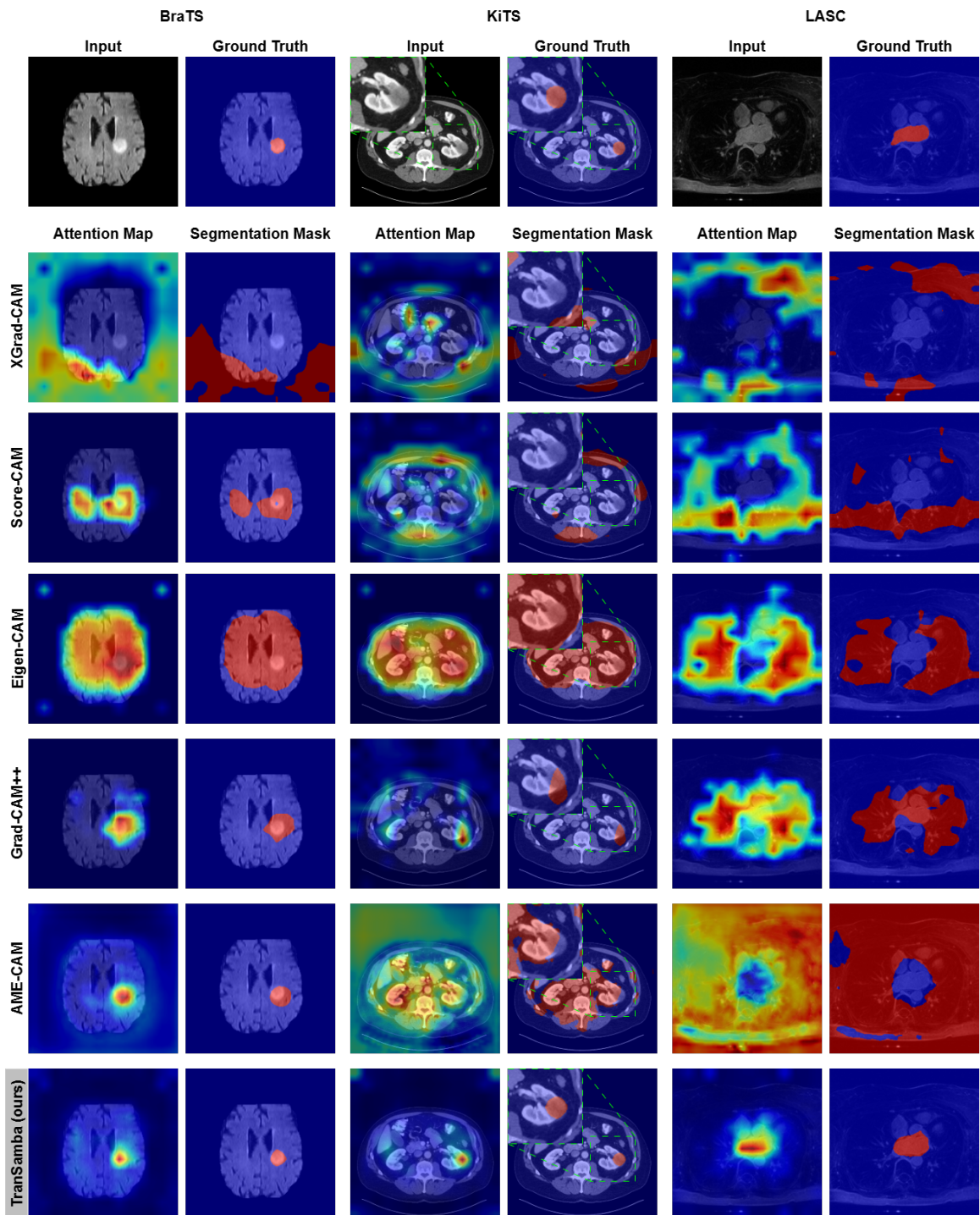
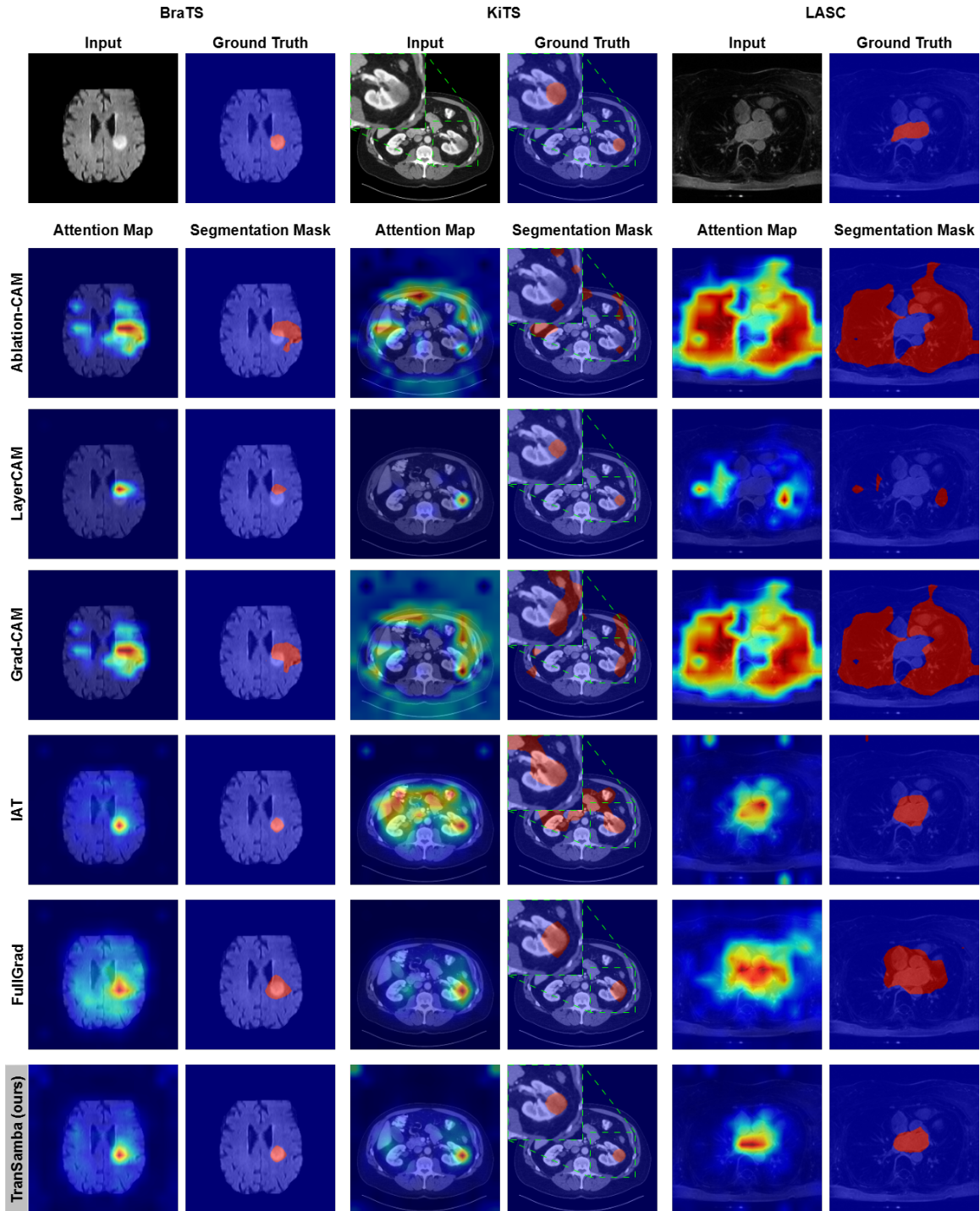Figure 6: Comprehensive qualitative comparison with SOTA, part A.

Figure 7: Comprehensive qualitative comparison with SOTA, part B.