Hybrid Transformer-Mamba for Weakly Supervised Volumetric Medical Segmentation

Yiheng Lyu[1,2(✉)], Lian Xu[1], Coen Arrow[1,2,3], Mohammed Bennamoun[1], Farid Boussaid[1], and Girish Dwivedi[1,2,4,5]

[1] University of Western Australia, Crawley WA 6009, Australia
yiheng.lyu@uwa.edu.au
[2] Harry Perkins Institute of Medical Research, Murdoch WA 6150, Australia
[3] National Imaging Facility, Nedlands WA 6009, Australia
[4] Fiona Stanley Hospital, Murdoch WA 6150, Australia
[5] Victor Chang Cardiac Research Institute, Darlinghurst NSW 2010, Australia

Abstract. Weakly supervised segmentation enables model training from plane-level labels. Existing methods often rely on 2D encoders, neglecting the volumetric nature of medical data. We propose TranSamba, a hybrid Transformer-Mamba architecture designed to capture 3D context via cross-plane modeling. TranSamba augments a Vision Transformer backbone with Cross-Plane Mamba blocks, leveraging linear-time modeling for efficient information exchange across neighboring planes. This exchange improves in-plane self-attention and subsequent attention maps for object localization. TranSamba maintains linear time complexity and constant space complexity with respect to the input volume depth. Extensive experiments on three datasets covering diverse modalities and pathologies show that TranSamba achieves state-of-the-art performance, demonstrating the generalizable efficacy of cross-plane modeling. Code is available at: `https://github.com/YihengLyu/TranSamba`.

Keywords: Hybrid Mamba · Weakly supervised medical segmentation.

1 Introduction

Segmentation models for volumetric medical images such as computed tomography (CT) and magnetic resonance imaging (MRI) are typically trained from voxel-level labels, which are laborious to obtain. Weakly supervised segmentation trains models from labor-efficient slice-level labels [7,12,25,32], which contain no localization information. Class activation mapping (CAM) [46] extracted from trained classification models can be used for coarse object localization. However, domain-specific characteristics introduce challenges for CAM-based methods. Existing strategies to tackle these challenges can be broadly categorized as: **1.** multi-scale feature aggregation [7,12,27]; **2.** addressing boundary and size characteristics [3,8,25,45]; **3.** self-supervised regularization [20,25,30,35]; **4.** incorporating target-specific constraints [5,8,9,21].

The volumetric nature remains underexplored when training segmentation models from slice-level labels. 2D encoders are typically used to make predictions from individual slices that align directly with their labels. The Vision Transformer (ViT) [10] is a popular 2D encoder; its class-to-patch (C2P) attention, the pairwise self-attention (SA) [37] between class and patch tokens, is effective for weakly supervised segmentation [43,44]. We term C2P attention within a single slice as **in-plane** SA, which is complemented by **cross-plane** interactions across neighboring slices to capture the volumetric nature. However, cross-plane SA is computationally expensive, scaling quadratically with N planes.

Built on structured state space models (SSMs), the Mamba architecture [14] achieves linear-time scaling with sequence length. Mamba-based vision backbones [23,47] have emerged for computational efficiency; in medical segmentation, Mamba applications have succeeded under fully [26,41] and weakly supervised [11,29] settings. Furthermore, Mamba-Transformer hybrids [15,22] are proposed alongside a novel pre-training strategy [24] for hybrid architectures.

To leverage linear-time cross-plane modeling via Mamba and the efficacy of C2P attention, we propose TranSamba, a hybrid **TranS**former-M**amba** architecture for weakly supervised volumetric medical segmentation. We place a novel Cross-Plane Mamba (CPM) block before the Transformer block in every layer of a ViT encoder. The CPM blocks allow efficient information exchange between patch tokens across neighboring planes, enriching information encoding for each token; the Transformer blocks produce C2P attention maps via in-plane SA, which benefits from this enrichment. Without any of the aforementioned strategies, TranSamba achieves improved object localization via straightforward model training. Our **contributions** are: **1.** We propose TranSamba, a hybrid architecture combining Transformer for effective in-plane SA with Mamba for efficient cross-plane modeling. **2.** The time complexity of TranSamba scales linearly with the number of planes; for batch processing, its space complexity remains constant with the number of planes. **3.** TranSamba achieves state-of-the-art (SOTA) performance on three distinct datasets, underscoring the generalizable effectiveness of cross-plane modeling for weakly supervised volumetric medical segmentation.

2 TranSamba

Fig. 1(a) shows the TranSamba overview. The encoder comprises L hybrid layers, each consisting of a CPM block in series with a Transformer block. An input volume comprises N planes, each 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; M and D denote the patch token number and the embedding dimension, respectively. In the l-th layer, the CPM input is obtained by stacking the $(l-1)$-th layer output sequences, and cross-plane modeling is achieved with the CPM block:

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

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

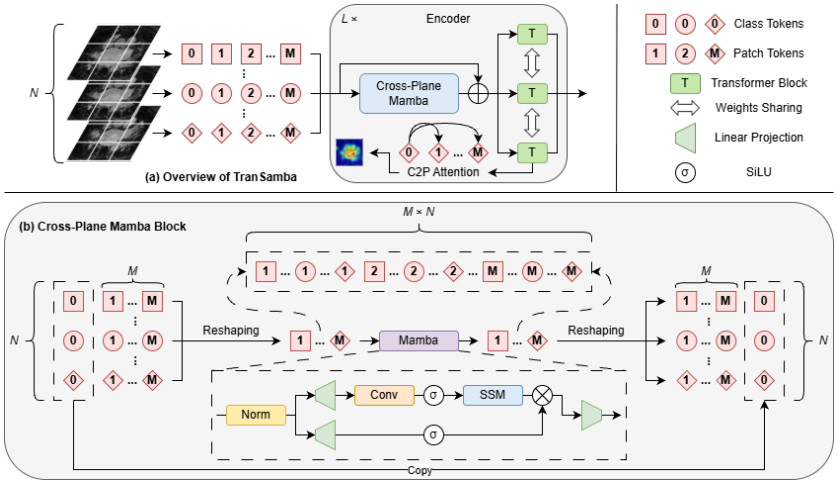

Fig. 1. (a) The encoder comprises L hybrid layers of CPM and Transformer blocks. An input volume contains N planes of $1 + M$ tokens. The N sequences are reshaped into a single sequence for cross-plane modeling with Mamba and reshaped to N sequences for parallelized in-plane SA. For inference, attention maps are generated from the C2P attention of the L Transformer blocks. **(b)** The $N \times M$ patch tokens are separated from class tokens and reshaped into a single sequence of length $M \times N$, where neighboring tokens are from different planes. Cross-plane modeling is achieved by learning interactions between the $M \times N$ tokens via Mamba. The Mamba output is reshaped into N sequences and concatenated with class tokens.

The CPM output is reshaped back to N sequences for parallelized in-plane SA:

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

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

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

During training, an output sequence $\mathbf{T}_L^{'}$ comprises a class token $\mathbf{T}_L^{class} \in \mathbb{R}^{1\times D}$ and M patch tokens $\mathbf{T}_L^{patch} \in \mathbb{R}^{M\times D}$. A global average pooling (GAP) layer processes $\mathbf{T}_L^{class}$: $\hat{y}^{class} = \text{GAP}(\mathbf{T}_L^{class})$; a convolutional layer followed by a global weighted ranking pooling (GWRP) [19] layer processes $\mathbf{T}_L^{patch}$: $\hat{y}^{patch} = \text{GWRP}(\text{Conv}(\mathbf{T}_L^{patch}))$. The training loss sums two cross-entropy (CE) terms: $\mathcal{L} = \text{CE}(\hat{y}^{class}, y) + \text{CE}(\hat{y}^{patch}, y)$. During inference, attention maps are generated by aggregating the C2P attention from all L Transformer blocks:

$$\mathbf{A}^{\text{C2P}} = \sum_{l=1}^{L} \mathbf{A}_l[1, 2:] \tag{5}$$

Table 1. Computational complexity comparison. Shaded cells represent TranSamba.

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

2.1 Cross-Plane Mamba

Fig. 1(b) shows a CPM block comprising a Mamba [14] block and two reshaping operations. The CPM input $\mathbf{V} \in \mathbb{R}^{N\times(1+M)\times D}$ comprises N sequences, each containing one class token for global plane representation and M patch tokens representing local patches. Cross-plane modeling with Mamba is achieved by efficiently learning interactions between the $M \times N$ patch tokens in linear time; cross-plane SA significantly increases computational complexity (Sec. 2.2).

The N class tokens are separated and excluded from the Mamba block. The $N \times M$ patch tokens are reshaped into a single sequence $\mathbf{x} \in \mathbb{R}^{(M\times N)\times D}$; neighboring patch tokens are from different planes, thereby prioritizing their interactions in cross-plane modeling with Mamba: $\mathbf{x}^{'} = \text{Mamba}(\mathbf{x})$. The Mamba output sequence $\mathbf{x}^{'}$ is reshaped into N parallel sequences of M patch tokens, each concatenated with its class token.

2.2 Computational Complexity Analysis

Given a sequence $\mathbf{T} \in \mathbb{R}^{1\times L\times D}$, the time complexity of global SA is $4LD^2 + 2L^2D$; with default expanded state dimension ($2D$) and fixed parameter (16), the time complexity of global SSM is $128LD$ [47]. For a volume with N planes, Tab. 1 summarizes the effective sequence length, batch size, and computational complexity of in-plane and cross-plane modeling. With the series arrangement of CPM and Transformer blocks, the time complexity of TranSamba is $128MND + 4MD^2 + 2M^2D$ and scales linearly with N. Replacing SSM with SA for cross-plane modeling increases the time complexity to $4MND^2 + 2M^2N^2D + 4MD^2 + 2M^2D$, scaling quadratically with N. For a batch of $\frac{B}{N}$ volumes each with N planes, the effective batch sizes for in-plane and cross-plane modeling are B and $\frac{B}{N}$. The space complexity of TranSamba is $B(4MD^2 + 2M^2D) + \frac{B}{N}(128MND)$, independent of N. Replacing SSM with SA for cross-plane modeling increases the space complexity to $B(4MD^2 + 2M^2D) + \frac{B}{N}(4MND^2 + 2M^2N^2D)$, which grows with N.

3 Experiments

We formulate the task as binary segmentation; while important, multi-class segmentation requires techniques orthogonal to architectural novelty, such as modeling co-occurrence [8] or inter-class [5,9,43,44] or inter-channel [20] correlations.

Datasets TranSamba is evaluated on three datasets. 1. Brain Tumor Segmentation (BraTS) [1,2,28]: using the T2 fluid-attenuated inversion recovery sequence for whole tumor segmentation; 2. Kidney Tumor Segmentation (KiTS) 2023 [16]: targeting the kidney tumor; 3. Left Atrium Segmentation Challenge (LASC) [42]: targeting the LA cavity. BraTS, KiTS, and LASC comprise 484, 489, and 154 volumetric studies, respectively; 100 studies per dataset are hidden for testing; the remaining studies use an 80/20 training/validation split.

Evaluation The ablation studies and SOTA comparisons are conducted on our validation and testing sets, respectively. The 3D mean Dice similarity coefficient (DSC) evaluates overall study-level segmentation performance; the 2D mean intersection-over-union (IoU) evaluates plane-level localization map quality. The 95% Hausdorff distance (HD95) is sensitive to outliers under weakly supervised settings and is only used for variants with close DSC/IoU.

Implementation Our ViT backbone is built on DeiT-S [36]; CPM blocks employ vanilla Mamba [14] blocks. For training, a batch comprises 16 randomly sampled volumes, each with 16 contiguous planes; all variants are trained for 100 epochs using only plane-level labels. We use the AdamW optimizer (initial learning rate 5×10^{-4}, weight decay 0.05) and a cosine decay schedule with a 5-epoch warmup. For inference, attention maps are generated in a single stage. Training and inference use a single 40GB NVIDIA A100 GPU.

3.1 Ablation Studies

We ablate **1.** cross-plane modeling effectiveness **2.** varying plane numbers, and **3.** hybrid layer design.

Effectiveness of Cross-Plane Modeling TranSamba is compared with no-Mamba variants; Tab. 2(a,b) presents the results. TranSamba (V3) outperforms V1, a ViT-only baseline, by large margins. To demonstrate this improvement is not from increased complexity, we introduce V2, also comprising alternating Transformer and in-plane Mamba blocks equivalent to the forward branch in Vision Mamba (ViM) [47]. With comparable complexity to V3, V2 yields much lower improvement over V1, indicating the performance improvement of V3 primarily stems from cross-plane modeling. Fig. 2(a) shows a qualitative comparison of V1–V3; V3 achieves the most conformal segmentation. For completeness, we report Mamba-only variant performance: V4 uses ViM-like bidirectional SSMs for in-plane modeling; V5 further augments V4 with CPM blocks; object localization relies on patch token-based CAM generation. Compared to V1–V3, V4 and V5 achieve substantially inferior performance, justifying using SA for in-plane modeling despite higher computational complexity; we acknowledge the lower speed of V4 and V5 results from implementation inefficiency. Nonetheless, V5 outperforms V4 by incorporating cross-plane modeling.

Table 2. Ablation studies. Shaded rows indicate the default configuration of TranSamba (V3, 16 planes, CrossIn). CPM: cross-plane Mamba; IPM: in-plane Mamba; IPT: in-plane Transformer; Uni-: unidirectional; Bi-: bidirectional.

(a) Complexity comparison.

Variant	CPM	IPM	IPT	Speed (FPS)	Memory (GB)	#params. (M)
V1	-	-	✓	3401.4	13.0	21.7
V2	-	Uni-	✓	1764.4	17.6	33.3
V3	✓	-	✓	1777.7	17.3	33.3
V4	-	Bi-	-	1613.1	9.2	23.6
V5	✓	Bi-	-	1126.6	13.5	35.2

(b) Effectiveness of cross-plane modeling.

Variant	BraTS		KiTS		LASC	
	DSC (%)	IoU (%)	DSC (%)	IoU (%)	DSC (%)	IoU (%)
V1	40.7	28.8	15.8	20.1	33.5	20.4
V2	41.9	28.2	19.5	24.7	38.1	24.3
V3	**60.7**	**44.2**	**25.1**	**24.9**	**45.9**	**29.3**
V4	32.6	18.3	5.8	5.3	0.3	0.1
V5	36.7	19.5	6.0	5.5	1.4	0.6

(c) Number of planes.

#planes	BraTS		KiTS		LASC	
	DSC (%)	IoU (%)	DSC (%)	IoU (%)	DSC (%)	IoU (%)
128	57.6	43.6	N/A	N/A	N/A	N/A
64	55.1	41.8	20.0	17.4	N/A	N/A
32	53.7	40.8	22.3	**27.7**	40.8	24.9
16	**60.7**	**44.2**	**25.1**	24.9	**45.9**	**29.3**

(d) Hybrid layer design.

Design	BraTS		KiTS		LASC	
	DSC (%)	IoU (%)	DSC (%)	IoU (%)	DSC (%)	IoU (%)
Parallel	60.8	42.0	23.9	24.4	**47.6**	**30.9**
InCross	**64.4**	**44.2**	24.6	**25.2**	42.0	26.2
CrossIn	60.7	**44.2**	**25.1**	24.9	45.9	29.3

Number of Planes We evaluate performance across varying plane numbers: the minimum is 16; the maximum cannot exceed the lowest dataset plane count (155/71/44 for BraTS/KiTS/LASC). As shown in Tab. 2(c), modeling across 16 planes achieves the highest performance. The target objects are continuous structures spanning a limited number of planes; hence, local cross-plane modeling is more beneficial than long-range modeling.

Hybrid Layer Design Regardless of the design, cross-plane modeling consistently improves the ViT baseline. TranSamba adopts the CrossIn design, which achieves the tied-highest ranking of 1.83 (Tab. 2(d)) and the lowest HD95 on KiTS and second-lowest HD95 on BraTS and LASC (BraTS/KiTS/LASC: *4.6*/**15.2**/*3.2*; Parallel: 4.8/15.5/**3.1**, InCross: **3.8**/*15.4*/3.6). A CrossIn layer places the CPM block before the Transformer block, ensuring C2P attention extraction follows cross-plane modeling, thereby maximizing map quality improvement.

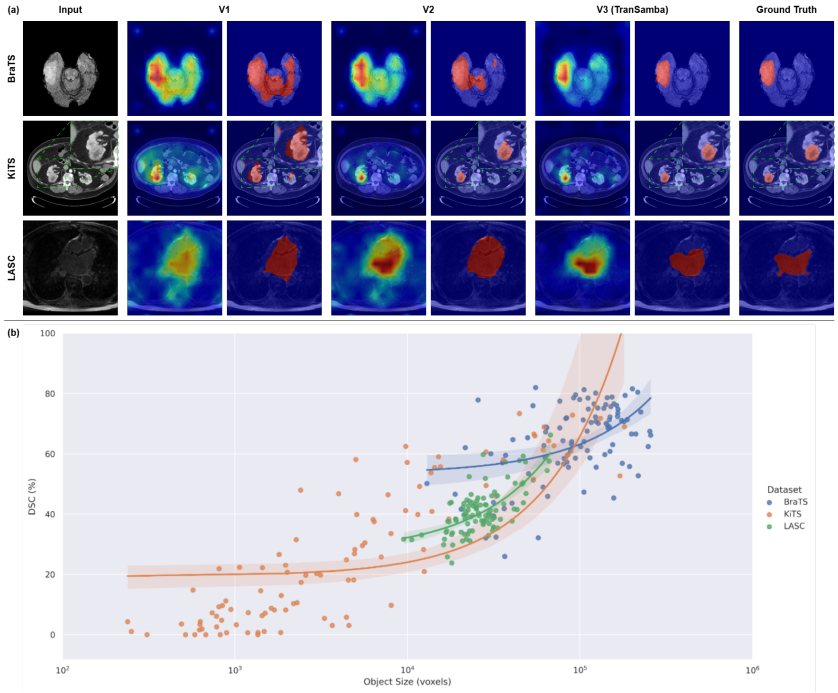

Fig. 2. (a) Qualitative comparison of V1–V3 (Tab. 2(a)) on the validation set. **(b)** Size-DSC correlation of TranSamba on the testing set.

3.2 Comparison with State-of-the-Art

Tab. 3 compares TranSamba with SOTA methods based on image-/plane-level labels. Most methods achieve higher performance on BraTS than the other datasets. While the BraTS background is normal brain tissue, KiTS and LASC backgrounds are noisy cavities; LASC also has a small training set (43 volumetric studies). TranSamba achieves the highest performance across all three datasets, outperforming the second-highest performer of each dataset by mean margins of 5.3% and 2.4% in DSC and IoU. Notably, AME-CAM [7] and UM-CAM [12] perform dynamic multi-scale feature aggregation, while IAT [25] adopts focal loss and self-supervision for intra-class heterogeneity. These strategies are effective on BraTS; however, performance drops substantially on KiTS and LASC. In contrast, TranSamba demonstrates robustness across tasks, indicating the generalizable effectiveness of cross-plane modeling.

Object Size Analysis To evaluate model sensitivity to object size; studies in the testing set are assigned to four subgroups according to target object size (in voxels): tiny, small, large, and huge (Tab. 4). Fig. 2(b) shows the size distribution and size-DSC correlation of TranSamba. BraTS and LASC objects have a more clustered distribution, while KiTS size distribution spans three orders of magnitude. Most KiTS objects are much smaller, resulting in the lower absolute

Table 3. Quantitative comparison with state-of-the-art methods on the testing set. Best and second-best results are highlighted in bold and italics, respectively. *: Hybrid backbone built on DeiT-S and vanilla Mamba.

Method	Backbone	BraTS		KiTS		LASC	
		DSC (%)	IoU (%)	DSC (%)	IoU (%)	DSC (%)	IoU (%)
Grad-CAM [33]	DeiT-S	55.2	36.6	10.7	15.0	6.0	2.8
Grad-CAM++ [4]	DeiT-S	45.9	28.3	3.8	5.6	8.8	5.9
FullGrad [34]	DeiT-S	33.2	19.8	20.5	*25.1*	*37.7*	*23.1*
Ablation-CAM [31]	DeiT-S	52.5	34.8	7.4	9.4	5.6	2.6
Score-CAM [39]	DeiT-S	32.4	21.9	0.8	0.8	4.3	2.0
SEAM [40]	ResNet-38	20.8	8.5	7.7	6.1	9.5	5.2
LayerCAM [17]	DeiT-S	29.4	17.8	*21.4*	17.3	0.3	0.2
TS-CAM [13]	DeiT-S	50.6	*42.8*	7.8	20.9	16.5	13.5
SIPE [6]	ResNet-50	19.9	10.5	3.3	2.7	4.0	1.9
AME-CAM [7]	ResNet-18	*57.0*	42.2	3.5	2.3	1.1	0.6
UM-CAM [12]	VGG-16	51.4	35.5	13.2	17.1	24.2	15.3
IAT [25]	DeiT-S	52.8	37.1	10.7	11.4	36.1	21.9
TranSamba	DeiT-S*	**64.1**	**46.8**	**26.8**	**25.8**	**41.0**	**25.6**

DSC on KiTS. Within each dataset, TranSamba performance exhibits a positive correlation with object size. Tab. 4 shows performance by subgroup; TranSamba also achieves the highest DSC for each subgroup, with the exception of the huge subgroup of KiTS (FullGrad: 66.1%). The size-DSC correlation indicates a limitation of TranSamba: performance drop on small objects, a well-known issue for medical segmentation [38].

Future Work TranSamba can be extended along two distinct avenues. First, having scoped this study to binary tasks to isolate the cross-plane modeling contribution, future extensions will incorporate multiple class tokens [43,44] to scale the architecture to multi-class scenarios. This will facilitate direct comparisons with methods utilizing target-specific loss functions [5,9]. Second, while TranSamba yields consistent improvements over baselines, its absolute performance can be further elevated for clinical deployment by integration with the Segment Anything Model (SAM) [18] family. Specifically, localization maps produced by TranSamba can serve as automated input prompts for SAM, eliminating manual interaction while boosting final segmentation accuracy to meet stringent clinical requirements.

4 Conclusion

This paper introduces TranSamba, a hybrid Transformer-Mamba architecture to bridge 3D modeling and the reliance on 2D encoders in weakly supervised volumetric medical segmentation. The integration of CPM blocks with a ViT encoder enables efficient cross-plane modeling, which improves in-plane SA for object localization. TranSamba maintains linear time complexity and constant space complexity with the input volume depth, and achieves SOTA performance on three datasets, demonstrating the generalizable effectiveness of cross-plane modeling across diverse volumetric medical segmentation tasks. Future work

Table 4. Quantitative results of TranSamba grouped by object size on the testing set.

Subgroup	Size (Voxels)	BraTS		KiTS		LASC	
		#studies	DSC (%)	#studies	DSC (%)	#studies	DSC (%)
Tiny	$< 10^3$	0	N/A	23	5.1	0	N/A
Small	$[10^3, 10^4)$	0	N/A	50	21.3	1	31.7
Large	$[10^4, 10^5)$	47	58.3	24	54.3	99	41.1
Huge	$\geq 10^5$	53	69.2	3	64.5	0	N/A

could extend TranSamba to multi-class scenarios and integrate TranSamba with the SAM family for more clinically oriented applications.

Acknowledgments. This work is supported by the University of Western Australia scholarships, the Australian Research Council grants (DE260101852, DP260101891), and the Medical Research Future Fund (NCRI000208).

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

References

1. Bakas, S., et al.: Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. SData (2017)
2. Bakas, S., et al.: Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. arXiv (2018)
3. Chang, R., et al.: Weakly-supervised ultrasound video segmentation with minimal annotations. In: MICCAI (2021)
4. Chattopadhay, A., et al.: Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: WACV (2018)
5. Chen, H., et al.: Ws-mtst: Weakly supervised multi-label brain tumor segmentation with transformers. JBHI (2023)
6. Chen, Q., et al.: Self-supervised image-specific prototype exploration for weakly supervised semantic segmentation. In: CVPR (2022)
7. Chen, Y.J., et al.: Ame-cam: Attentive multiple-exit cam for weakly supervised segmentation on mri brain tumor. In: MICCAI (2023)
8. Chen, Z., et al.: C-cam: Causal cam for weakly supervised semantic segmentation on medical image. In: CVPR (2022)
9. Dhamale, V., Sundaresan, V.: Inter-class separability loss for weakly supervised mutually exclusive multiclass segmentation of brain tumor lesions. In: MICCAI (2025)
10. Dosovitskiy, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv (2020)
11. Fan, J., et al.: Pathmamba: Weakly supervised state space model for multi-class segmentation of pathology images. In: MICCAI (2024)
12. Fu, J., et al.: Um-cam: Uncertainty-weighted multi-resolution class activation maps for weakly-supervised fetal brain segmentation. In: MICCAI (2023)
13. Gao, W., et al.: Ts-cam: Token semantic coupled attention map for weakly supervised object localization. In: ICCV (2021)

14. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. In: COLM (2024)
15. Hatamizadeh, A., Kautz, J.: Mambavision: A hybrid mamba-transformer vision backbone. In: CVPR (2025)
16. Heller, N., et al.: The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. MedIA (2021)
17. Jiang, P.T., et al.: Layercam: Exploring hierarchical class activation maps for localization. TIP (2021)
18. Kirillov, A., et al.: Segment anything. In: ICCV (2023)
19. Kolesnikov, A., Lampert, C.H.: Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In: ECCV (2016)
20. Kuang, Z., et al.: Cluster-re-supervision: Bridging the gap between image-level and pixel-wise labels for weakly supervised medical image segmentation. JBHI (2023)
21. Li, Y., et al.: Deep weakly-supervised breast tumor segmentation in ultrasound images with explicit anatomical constraints. MedIA (2022)
22. Lieber, O., et al.: Jamba: A hybrid transformer-mamba language model. arXiv (2024)
23. Liu, Y., et al.: Vmamba: Visual state space model. NeurIPS (2024)
24. Liu, Y., Yi, L.: Map: Unleashing hybrid mamba-transformer vision backbone's potential with masked autoregressive pretraining. In: CVPR (2025)
25. Lyu, Y., et al.: Importance-aware transformer: Addressing intra-class heterogeneity in weakly supervised brain tumor segmentation. In: DICTA (2024)
26. Ma, J., et al.: U-mamba: Enhancing long-range dependency for biomedical image segmentation. arXiv (2024)
27. Ma, X., et al.: Ms-cam: Multi-scale class activation maps for weakly-supervised segmentation of geographic atrophy lesions in sd-oct images. JBHI (2020)
28. Menze, B.H., et al.: The multimodal brain tumor image segmentation benchmark (brats). TMI (2014)
29. Pan, Z., et al.: Mamba-based weakly supervised medical image segmentation with cross-modal textual information. In: MICCAI (2025)
30. Patel, G., Dolz, J.: Weakly supervised segmentation with cross-modality equivariant constraints. MedIA (2022)
31. Ramaswamy, H.G., et al.: Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. In: WACV (2020)
32. Schmidt-Mengin, M., et al.: Tonno: Tomographic reconstruction of a neural network's output for weakly supervised segmentation of 3d medical images. In: CVPR (2024)
33. Selvaraju, R.R., et al.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: ICCV (2017)
34. Srinivas, S., Fleuret, F.: Full-gradient representation for neural network visualization. NeurIPS (2019)
35. Tang, W., et al.: M-seam-nam: multi-instance self-supervised equivalent attention mechanism with neighborhood affinity module for double weakly supervised segmentation of covid-19. In: MICCAI (2021)
36. Touvron, H., et al.: Training data-efficient image transformers & distillation through attention. In: ICML (2021)
37. Vaswani, A., et al.: Attention is all you need. NeurIPS (2017)
38. Wang, G., et al.: S^3-mamba: Small-size-sensitive mamba for lesion segmentation. In: AAAI (2025)
39. Wang, H., et al.: Score-cam: Score-weighted visual explanations for convolutional neural networks. In: CVPRW (2020)

40. Wang, Y., et al.: Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In: CVPR (2020)
41. Xing, Z., et al.: Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. In: MICCAI (2024)
42. Xiong, Z., et al.: A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging. MedIA (2021)
43. Xu, L., et al.: Multi-class token transformer for weakly supervised semantic segmentation. In: CVPR (2022)
44. Xu, L., et al.: Mctformer+: Multi-class token transformer for weakly supervised semantic segmentation. TPAMI (2024)
45. Yang, J., et al.: Anomaly-guided weakly supervised lesion segmentation on retinal oct images. MedIA (2024)
46. Zhou, B., et al.: Learning deep features for discriminative localization. In: CVPR (2016)
47. Zhu, L., et al.: Vision mamba: Efficient visual representation learning with bidirectional state space model. arXiv (2024)