Mamba-CAM: Mamba-Driven Class Activation Map Learning for 3D Medical Weakly-Supervised Segmentation

Xiaofeng Han[2,1†], Ganggang Liu[3†], Shunpeng Chen[4†], Mingming Yu[5], Duzhen Zhang[1,2], Xinchen Li[6], Changwei Wang[7,8], Weiliang Meng[1,2*], Xiaopeng Zhang[1,2*]

[1]MAIS, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China
[3]Shanghai Institute of Aerospace Technical Foundation, Shanghai 200000, China
[4]Beijing University of Posts and Telecommunications, Beijing 100876, China
[5]Beihang University, Beijing 100191, China
[6]Tongji University, Shanghai 200092, China
[7]Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center, Qilu University of Technology, Jinan 250013, China
[8]Shandong Provincial Key Laboratory of Computing Power Internet and Service Computing, Shandong Fundamental Research Center for Computer Science, Jinan 250013, China

***Abstract*—Weakly supervised medical image segmentation methods commonly rely on convolutional neural networks (CNNs) and class activation maps (CAMs). However, the inherent locality of CNNs often leads to suboptimal CAM quality and degraded segmentation performance. To overcome this limitation, we propose Mamba-CAM, a novel weakly supervised segmentation framework that integrates CAMs with the Mamba model. Leveraging State Space Models (SSMs) for efficient long-range contextual modeling, our Mamba-CAM generates more reliable pseudo-labels for training. To further enhance interpretability and detail preservation, we design a detail-aware visual state space (DVSS) block. Moreover, a correspondence learning mechanism is provided to refine pseudo-labels and improve feature representation. To the best of our knowledge, this is the first work to integrate Mamba with CAMs for medical image segmentation. Extensive experiments on the BraTS dataset demonstrate that our Mamba-CAM achieves state-of-the-art performance, highlighting its potential to advance medical image analysis while reducing annotation demands.**

***Index Terms*—medical image segmentation, weakly-supervised learning, CAM, Mamba**

I. Introduction

With the rapid development of deep learning, the medical imaging field has undergone profound transformation. However, medical segmentation models still largely rely on high-quality pixel-level annotations, which are expensive and time-consuming to obtain. To alleviate this bottleneck, weakly supervised segmentation has attracted increasing attention: it learns from coarse-grained or sparse annotations, thereby improving the practicality and scalability of medical segmentation in real clinical settings.

† These authors contributed equally.
* Corresponding author:weiliang.meng@ia.ac.cn;xiaopeng.zhang@ia.ac.cn

Most existing weakly supervised segmentation methods are built upon CNNs and CAMs. Yet, medical image segmentation is inherently a dense and global task. CNNs are constrained by their local receptive fields, while CAMs are often coarse and incomplete, leading to limited performance in complex and heterogeneous medical imaging scenarios. Therefore, the central challenge lies in effectively capturing global context in medical imagery and constructing more refined and reliable CAMs to provide stronger supervisory signals for segmentation.

Recent studies [1]–[3] have explored cost-effective supervision across various weak annotation types, ranging from bounding boxes to scribbles/points. Lee et al. [4] introduced bounding-box attribution maps (BBAM) to generate pseudo masks, and related works [5]–[7] further demonstrate that scribble/point supervision can reduce labeling effort, although some manual input is still required. Moreover, dense segmentation using only image-level class labels has emerged as a promising direction [3], [8]–[10]. However, the complex anatomy, ambiguous boundaries, and cross-modality appearance variations in medical images often degrade CAM quality, which in turn undermines the reliability of pseudo-label supervision.

To learn robust representations under these challenges, modeling global context becomes particularly important. Mamba [11], built on SSMs [12], models long-range dependencies with linear complexity $O(n)$ and is more efficient than the quadratic $O(n^2)$ complexity of Transformers [13]. V-Mamba [14] further extends this paradigm to vision tasks and shows strong performance on high-resolution images. In medical segmentation, SegMamba [15] and Mamba-UNet [16] also validate its advantages in long-range dependency modeling and local-global representation fusion, suggesting that

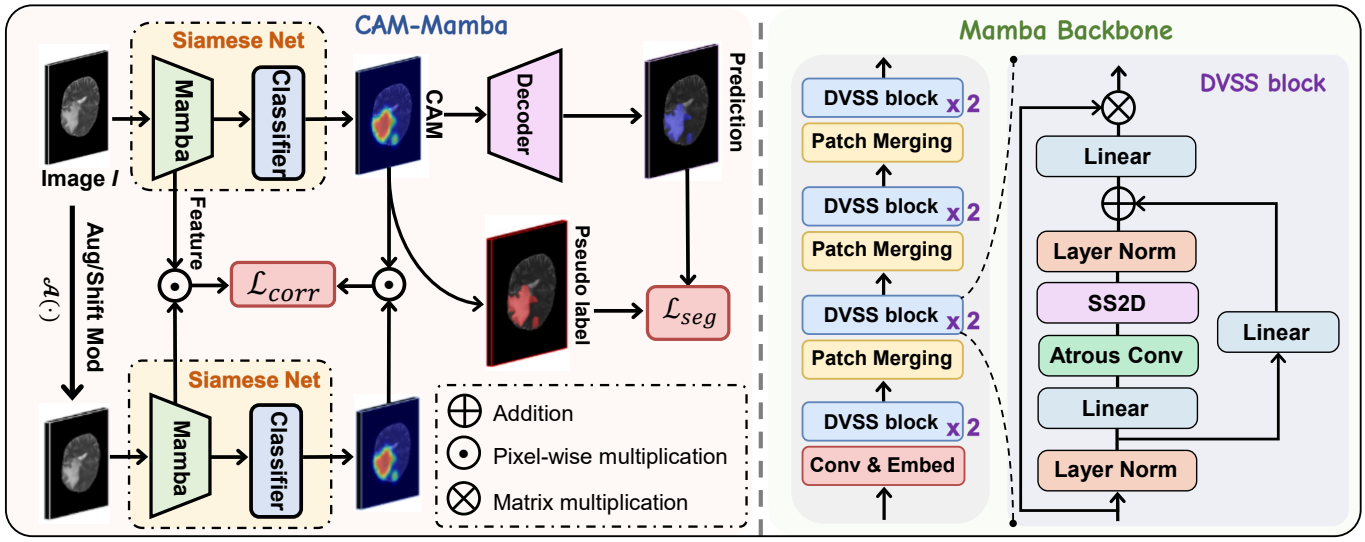

Fig. 1. Overview of our Mamba-CAM. Left: Input images are used to generate CAMs for pseudo-label creation and decoder supervision with a segmentation loss function $\mathcal{L}_{seg}$. Transformations $\mathcal{A}(\cdot)$, through data augmentation or modality changes, produce varied CAMs and images, reinforcing Mamba feature robustness via consistent correspondence. Right: The enhanced Mamba backbone featuring the detail-aware visual state space block (DVSS) with atrous convolution to sharpen detail sensitivity for complex medical imaging tasks.

introducing Mamba into weakly supervised medical segmentation may enhance both contextual modeling and the quality of CAM-based pseudo labels.

In summary, the key to weakly supervised medical segmentation lies in global context modeling and CAM refinement, which are crucial for improving localization and pseudo-label reliability. To this end, we propose Mamba-CAM, which integrates Mamba into the CAM generation process to produce more complete and accurate CAMs, while potentially improving robustness across medical images of varying resolutions. In addition, we design a DVSS module to better capture fine-grained structures. In our framework, CAMs generated by the Mamba backbone serve as initial pseudo labels, which are then further refined through a correspondence learning mechanism, where consistency constraints strengthen feature learning and improve segmentation performance.

Our main contributions are summarized as follows:

- We propose Mamba-CAM, a novel framework that exploits Mamba's global context modeling to generate refined, well-structured pseudo-labels, thereby substantially improving segmentation quality. To the best of our knowledge, this is the first work to integrate Mamba with CAMs for weakly supervised medical image segmentation.
- We design a DVSS block, an enhanced extension of the Mamba backbone that strengthens fine-detail perception and is particularly important for learning from complex medical imagery.
- We provide a correspondence learning mechanism to refine CAM-derived pseudo-labels, improving feature consistency and robustness across diverse medical imaging modalities. Our method achieves state-of-the-art performance on widely used benchmarks such as the BraTS [17] dataset.

II. Method

A. Overview of our Mamba-CAM

Fig. 1 shows our Mamba-CAM framework. The Mamba backbone, integrated within a Siamese network, enhances feature extraction for medical images. The DVSS blocks employ a 2D selective scan mechanism to traverse the data in four directions, enabling the model to capture rich contextual information from multiple perspectives.

B. CAM-based Pseudo-label Initialization

CAM [18] is widely adopted in image-level weakly supervised semantic segmentation to localize class-discriminative regions and provide initial pseudo-labels. In our Mamba-CAM framework, CAM serves as the starting point for constructing pixel-wise supervision from image-level annotations.

Given an input image I, our classification backbone (Mamba-based encoder) extracts a set of feature maps. Let $f_k(x, y)$ denote the activation at spatial location (x, y) in the k-th feature map. We first apply Global Average Pooling (GAP) to obtain the pooled feature for each channel:

$$F_k = \frac{1}{H \times W} \sum_{x=1}^{H} \sum_{y=1}^{W} f_k(x, y), \tag{1}$$

where H and W are the height and width of the feature map. The pooled features are then fed into a fully connected (FC) classification layer, where w_k^c denotes the weight associated with channel k for class c. The class score (before softmax) is computed as:

$$S_c = \sum_k w_k^c F_k. \tag{2}$$

To obtain the CAM for class c, we project the class-specific weights back onto the spatial feature maps:

$$M_c(x, y) = \sum_k w_k^c f_k(x, y). \tag{3}$$

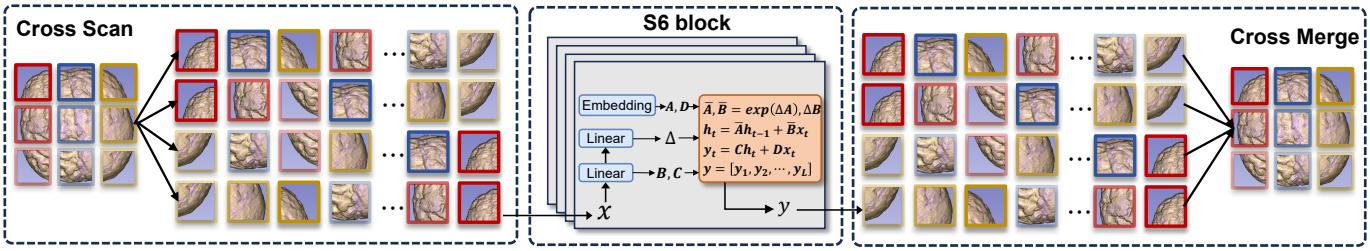

Fig. 2. SS2D with cross scan, S6 modeling, and cross merge. The feature map is scanned in multiple directions and encoded as sequences for the S6 block, enabling efficient long-range context modeling. The outputs are finally merged back to maintain structural consistency in the recovered spatial representation.

This yields a coarse localization map indicating the regions most relevant to class c. We normalize M_c and use it to initialize pseudo masks, which are subsequently refined by our Mamba-based correspondence learning for stronger and more reliable segmentation supervision.

C. Detail-aware VSS block

The spatial state sequence model [12] adopted by Mamba enables long-range dependency modeling with linear-time complexity, avoiding the quadratic cost of self-attention and thus being well-suited for dense prediction in high-resolution medical images. Inspired by V-Mamba [14], we design a *Detail-aware Visual State Space* (DVSS) block in our Mamba-CAM framework to strengthen the sensitivity to fine structures, which is crucial for weakly supervised medical segmentation where low contrast and subtle boundaries often dominate delineation quality.

Specifically, DVSS first applies LayerNorm to the input feature map and then splits it into two branches. The *semantic branch* employs lightweight linear projections with non-linear activations to preserve discriminative semantics. The *detail branch* injects a detail-enhancement prior via multi-rate atrous convolutions, capturing multi-scale context while preserving resolution, thereby improving edge/structure awareness under heterogeneous imaging conditions. The two branches are aggregated and fed into the 2D Selective Scan module [14] for global context modeling. SS2D follows a *cross-scan* $\rightarrow$ *S6 block* $\rightarrow$ *cross-merge* pipeline(as shown in Fig. 2): it first unfolds the 2D feature map into four 1D sequences scanned from different directions, then applies selective state-space modeling (S6) on each sequence to capture long-range dependencies, and finally merges the four outputs back to the original spatial layout, reconstructing the feature map while retaining multi-directional global context. After SS2D, both branches are normalized and fused via element-wise multiplication to effectively integrate fine-grained details with long-range context. The fused representation is further refined by a linear layer and added back to the original input through a residual connection, producing the final DVSS output.

D. Mamba-based Correspondence

We employ CAMs to generate initial pseudo-labels, with a focus on capturing correlations between CAMs from different images. Given two CAMs, $f_1 \in \mathbb{R}^{H_1 \times W_1 \times C}$ and $f_2 \in \mathbb{R}^{H_2 \times W_2 \times C}$, which share the same category dimension C but differ in spatial dimensions, we compute the feature correspondence matrix $M \in \mathbb{R}^{H_1 W_1 \times H_2 W_2}$ using the normalized dot product of their flattened representations:

$$\mathcal{M}(i,j) = \frac{\text{flatten}(f_1)(i) \cdot \text{flatten}(f_2)(j)}{\|\text{flatten}(f_1)(i)\| \|\text{flatten}(f_2)(j)\|} \tag{4}$$

As shown in Fig. 1, the encoder and decoder share weights in our Mamba-CAM framework. The Mamba feature maps for the original image I and its augmented counterpart $\mathcal{A}(I)$ are denoted as $s_1 \in \mathbb{R}^{H_1 \times W_1 \times C}$ and $s_2 \in \mathbb{R}^{H_2 \times W_2 \times C}$, respectively. The Mamba feature correspondence $\mathcal{S} \in \mathbb{R}^{H_1 W_1 \times H_2 W_2}$ is computed analogously to $\mathcal{M}$, based on s_1 and s_2.

This mechanism systematically captures and leverages spatial correlations between class activation maps, which is critical for improving object localization and segmentation performance. To further enhance the consistency and robustness of Mamba features, we introduce the correspondence loss $\mathcal{L}_{corr}$. Given the CAM correspondence matrix $\mathcal{M} \in \mathbb{R}^{H_1 W_1 \times H_2 W_2}$ and the Mamba feature correspondence matrix $\mathcal{S} \in \mathbb{R}^{H_1 W_1 \times H_2 W_2}$, the correspondence loss is defined as:

$$\mathcal{L}_{corr} = -\frac{1}{H_1 W_1 \times H_2 W_2} \sum_{i=1}^{H_1 W_1} \sum_{j=1}^{H_2 W_2} \mathcal{M}(i,j) \cdot \mathcal{S}(i,j) \tag{5}$$

E. Training Objective

To effectively supervise our Mamba-CAM, we define three objective functions. The first function $\mathcal{L}_{cls}$ trains the Mamba backbone together with the classification layer using only image-level labels. For a dataset with K classes, let $\bar{y}_c$ denote the model's predicted probability for class c, and $t_c \in \{0,1\}$ indicate the ground-truth label (1 for presence, 0 for absence). The classification loss is formulated as:

$$\mathcal{L}_{cls} = -\sum_{c=1}^{K} t_c \log(\bar{y}_c) \tag{6}$$

The second function is pixel-wise supervision $\mathcal{L}_{seg}$ for the segmentation task, which is defined as:

$$\mathcal{L}_{seg} = -\frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \sum_{c=1}^{K} t_c^{i,j} \log(\bar{y}_c^{i,j}) \tag{7}$$

where H and W denote the height and width of the image, and $t_c^{i,j}$ and $\bar{y}_c^{i,j}$ represent the ground-truth label and predicted probability for class c at pixel (i,j).

The third function is the overall training loss $\mathcal{L}$, which combines the above components, weighted by empirically chosen hyperparameters λ_1 and λ_2 (both set to 0.1):

$$\mathcal{L} = \lambda_1 \mathcal{L}seg + \lambda_2 \mathcal{L}_{corr} + \mathcal{L}_{cls} \tag{8}$$

This formulation enables the model to learn effectively from both image-level and pixel-wise supervision, while ensuring consistency between CAMs and Mamba features.

III. Experiments

A. Settings

Dataset. We evaluate our Mamba-CAM on the rain Tumor Segmentation (BraTS) dataset [17], which includes 2,000 patient cases. Each case consists of four 3D MRI modalities—T1, T1 with contrast enhancement (T1-CE), T2, and T2-FLAIR—along with a segmented ground-truth mask for validation. The dataset is split into training, validation, and test sets using an 8:1:1 ratio, comprising 5,802 positive and 1,073 negative samples. For preprocessing, the 3D volumes are sliced along the z-axis, resulting in $93,905$ 2D slices, following the approach of prior work [19]. Notably, the segmentation masks are used only for evaluation and are not employed during model training.

Implementation Details. Our Mamba-CAM is implemented in PyTorch, with the enhanced Mamba serving as the backbone network. For fair comparison, we pre-trained the network using SupCon [20] following the protocol in [19]. During fine-tuning, we used the AdamW optimizer, with the learning rate linearly increasing to 1×10^{-4} over the first 2,000 iterations, followed by a polynomial decay schedule. The warm-up and decay rates were set to 1×10^{-6} and 0.9, respectively.

For evaluation, Dice score and Intersection over Union (IoU) were used to measure segmentation quality, consistent with prior studies [19], [21], and 95% Hausdorff Distance (HD95) was reported to evaluate boundary precision.

B. Comparisons

We conduct a comprehensive benchmarking of state-of-the-art weakly supervised semantic segmentation methods, including AME-CAM [19], Grad-CAM [22], ScoreCAM [23], LFI-CAM [24], LayerCAM [25], and Swin-MIL [21], establishing a solid reference point for medical image segmentation. To provide deeper insight into performance, we further incorporate both fully supervised and unsupervised metrics from the C&F [26] framework. In addition, we include a comparative analysis against the fully supervised Optimized U-Net [27]. It is important to note that direct comparison across fully supervised, weakly supervised, and unsupervised methods is inherently limited due to their reliance on different levels of annotation.

Quantitative Comparison. Table I shows that our Mamba-CAM consistently outperforms all competing methods across the MRI modalities of the BraTS dataset. Notably, the strongest weakly supervised baseline, AME-CAM [19], achieves results closest to ours, highlighting the importance of capturing fine-grained structural details for accurate segmentation. These findings demonstrate that, under weak supervision, our proposed DVSS block effectively enhances detail perception, enabling more robust delineation of tumor boundaries and complex tissue structures. Compared with traditional CAM-based approaches such as Grad-CAM [22] and ScoreCAM [23], Mamba-CAM not only achieves higher Dice and IoU scores but also exhibits superior HD95 boundary accuracy, which confirms the effectiveness of combining Mamba's global context modeling with detail-aware enhancements.

In contrast, the unsupervised variant of C&F [26] performs poorly across all modalities, with consistently low Dice scores and extremely large HD95 errors, reflecting the limitations of relying solely on unsupervised learning signals. Although the supervised variant of C&F yields substantial improvements, it still lags behind our Mamba-CAM, suggesting that in the absence of dense supervision, the key to improved performance lies in effective global modeling and enhanced detail perception mechanisms.

Although the fully supervised Optimized U-Net [27] remains the strongest overall method, achieving the highest Dice scores and the lowest boundary errors across all modalities, our approach significantly narrows the performance gap between weakly and fully supervised methods. For example, on the T2-FLAIR modality, the Dice score of Mamba-CAM already approaches that of the fully supervised model. This result not only emphasizes the potential of weakly supervised segmentation in medical imaging, but also demonstrates that competitive performance can be achieved with substantially reduced annotation costs, offering a practical and promising solution for clinical applications.

Visual Comparison. As shown in Fig. 3, the visual comparison of segmentation results on the BraTS dataset highlights the clear advantages of our Mamba-CAM over existing weakly supervised methods. Grad-CAM [22] and LayerCAM [25] struggle to accurately localize tumor regions, often producing incomplete or diffuse predictions. Swin-MIL [21] generally provides reasonable segmentations but tends to overextend boundaries, introducing numerous false positives and reducing Dice and IoU scores. LFI-CAM [24] and AME-CAM [19] achieve closer alignment with ground truth in both size and location, yet remain limited in capturing fine structural details. In contrast, our Mamba-CAM not only accurately localizes tumors and estimates their size, but also better preserves morphological consistency through stronger detail perception. Furthermore, aided by the correspondence learning mechanism, Mamba-CAM remains robust on the challenging T2 modality and consistently outperforms other methods.

C. Ablation Study

We conduct ablation studies on the BraTS dataset [17] to systematically evaluate the contribution of each component in Mamba-CAM under comparable settings, using Dice↑/IoU↑/HD95↓ as metrics. Table II reports the results across Dice, IoU, and HD95 metrics for various module configurations. Starting from the Mamba baseline, adding

TABLE I

COMPARATIVE RESULTS OF SEGMENTATION METHODS ON THE BRATS DATASET USING MRI MODALITIES (T1, T1-CE, T2, T2-FLAIR). METHODS ARE GROUPED BY SUPERVISION TYPE: WEAKLY SUPERVISED SEGMENTATION (WSSS, †), UNSUPERVISED LEARNING (‡), AND FULLY SUPERVISED LEARNING (⋆). RESULTS ARE REPORTED AS MEAN ± STANDARD DEVIATION. THE BEST-PERFORMING WSSS METHODS ARE HIGHLIGHTED IN **BOLD**.

Modality	T1			T1-CE		
Method	Dice△	IoU△	HD95▽	Dice△	IoU△	HD95▽
Grad-CAM†	0.11±0.09	0.06±0.06	121.82±22.96	0.13±0.09	0.07±0.05	129.89±27.85
ScoreCAM†	0.30±0.13	0.18±0.09	60.30±14.11	0.40±0.19	0.27±0.16	46.83±22.09
LFI-CAM†	0.57±0.17	0.41±0.15	23.94±25.61	0.12±0.12	0.07±0.07	136.25±38.62
LayerCAM†	0.57±0.17	0.42±0.16	23.34±27.37	0.51±0.21	0.37±0.18	29.85±45.88
Swin-MIL†	0.48±0.17	0.33±0.15	46.47±30.41	0.46±0.17	0.31±0.14	47.00±22.82
AME-CAM†	0.63±0.12	0.47±0.12	21.81±18.22	0.70±0.10	0.54±0.11	18.130±12.34
Ours†	**0.67±0.11**	**0.51±0.10**	**20.91±10.78**	**0.73±0.12**	**0.60±0.09**	**15.51±10.40**
C&F‡	0.20±0.08	0.11±0.05	79.19±14.30	0.180±0.08	0.10±0.05	77.98±14.04
C&F⋆	0.57±0.10	0.42±0.19	29.02±20.88	0.25±0.10	0.14±0.07	130.62±9.88
Opt. U-net⋆	0.84±0.06	0.72±0.09	11.73±10.35	0.85±0.06	0.74±0.09	11.59±11.12
Modality	T2			T2-FLAIR		
Method	Dice△	IoU△	HD95▽	Dice△	IoU△	HD95▽
Grad-CAM†	0.05±0.06	0.03±0.03	141.03±23.11	0.15±0.08	0.08±0.05	110.03±23.31
ScoreCAM†	0.53±0.18	0.38±0.17	28.61±11.60	0.43±0.21	0.30±0.18	39.39±17.18
LFI-CAM†	0.67±0.17	0.53±0.19	18.17±10.48	0.16±0.19	0.10±0.14	125.75±45.58
LayerCAM†	0.62±0.18	0.48±0.17	23.98±44.32	0.65±0.21	0.52±0.21	22.06±33.96
Swin-MIL†	0.44±0.15	0.29±0.12	38.01±30.00	0.27±0.12	0.16±0.08	41.87±19.23
AME-CAM†	0.72±0.09	0.57±0.10	14.94±8.74	0.86±0.09	0.77±0.12	8.66±6.44
Ours†	**0.73±0.06**	**0.59±0.05**	**12.87±6.52**	**0.87±0.09**	**0.79±0.11**	**7.01±8.25**
C&F‡	0.23±0.09	0.13±0.06	76.26±13.19	0.10±0.05	0.20±0.17	75.65±14.21
C&F⋆	0.61±0.22	0.47±0.22	109.82±27.74	0.58±0.14	0.42±0.13	138.14±14.28
Opt. U-net⋆	0.88±0.06	0.80±0.10	8.35±9.13	0.91±0.06	0.85±0.09	8.09±11.88

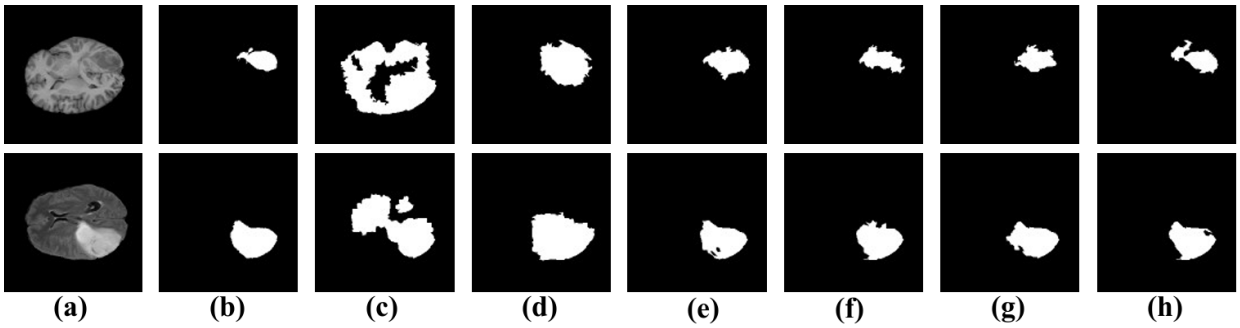

Fig. 3. Visual comparison of all methods on BraTS [17]. (a) Input Image. (b) Ground Truth. (c) Grad-CAM. (d) ScoreCAM. (e) LFI-CAM. (f) LayerCAM. (g) AME-CAM. (h) Mamba-CAM (ours).

TABLE II

ABLATION STUDIES OF OUR MAMBA-CAM ON BRATS [17].

Module Selection DVSS	$\mathcal{L}_{corr}$	Dice△	IoU△	HD95▽
		0.56±0.13	0.42±0.12	30.24±13
✔		0.63±0.11	0.46±0.11	23.21±12
	✔	0.64±0.11	0.47±0.11	21.35±12
✔	✔	**0.67±0.11**	**0.51±0.10**	**20.91±10.78**

DVSS increases Dice to 0.63; within its dual-branch design, the linear branch provides efficient token mixing, while the atrous-convolution branch introduces multi-rate receptive fields, enabling better capture of fine-grained structures while preserving global context. Further introducing the correspondence learning mechanism raises Dice to 0.64 by enforcing cross-sample/augmentation consistency to stabilize features and refine CAMs, thereby reducing spurious activations and improving region localization. Using both together yields cumulative gains: the full model reaches 0.67 Dice with higher IoU and lower HD95, indicating not only better overlap but also more accurate boundaries.

TABLE III

ABLATION STUDY OF THE DVSS STRUCTURE. WE COMPARE THE FULL DVSS WITH A SIMPLIFIED VARIANT WITHOUT THE MULTI-SCALE ATROUS CONVOLUTION BRANCH, SHOWING THE IMPORTANCE OF THIS COMPONENT FOR SEGMENTATION.

DVSS Configuration	**Dice ↑**	**IoU ↑**	**HD95 ↓**
w/o Atrous Convolution Branch	0.59 ± 0.12	0.44 ± 0.11	26.55 ± 12.81
Full DVSS (Ours)	**0.67 ± 0.11**	**0.51 ± 0.10**	**20.91 ± 10.78**

TABLE IV

COMPARATIVE RESULTS OF DIFFERENT BACKBONE ARCHITECTURES. MAMBA DEMONSTRATES SUPERIOR LONG-RANGE DEPENDENCY MODELING ABILITY IN MEDICAL IMAGE SEGMENTATION.

Backbone Architecture	**Dice ↑**	**IoU ↑**	**HD95 ↓**
ResNet-50 (CNN)	0.58 ±0.13	0.43 ±0.12	28.16 ±13.05
Swin-T (Transformer)	0.63 ±0.11	0.47 ±0.11	23.45 ±11.52
Mamba (Ours)	**0.67 ±0.11**	**0.51 ±0.10**	**20.91 ±10.78**

To further analyze DVSS (Table III), we remove the atrous-convolution branch and keep only the linear branch; this linear-only DVSS clearly degrades performance (Dice 0.59,

IoU 0.44, HD95 26.55), whereas the full DVSS achieves (Dice 0.67, IoU 0.51, HD95 20.91), demonstrating that the multi-scale context provided by atrous convolutions is indispensable—especially for depicting complex structures and fuzzy tumor margins. Finally, under the same framework (Table IV), using Mamba as the backbone attains 0.67/0.51/20.91, outperforming ResNet-50 (0.58/0.43/28.16) and Swin-T (0.63/0.47/23.45), highlighting the advantage of state-space modeling for capturing long-range dependencies and explaining the consistent IoU/HD95 improvements observed across our ablations.

IV. Conclusion

We propose Mamba-CAM, a novel integration of the Mamba model with CAMs, driving significant progress in weakly supervised medical image segmentation. By leveraging Mamba's capacity for modeling global context, our Mamba-CAM produces more reliable pseudo-labels, thereby improving segmentation performance under limited annotation. The designed DVSS block enhances the model's ability to capture fine-grained structural details in complex medical images, which is critical for accurate diagnosis. In addition, our correspondence learning mechanism refines pseudo-label quality and promotes consistent feature learning across modalities. These innovations achieve state-of-the-art results on the BraTS dataset, underscoring the effectiveness of our framework. Beyond advancing segmentation accuracy, this work opens promising directions for future research in medical imaging and the broader application of weakly supervised learning.

Acknowledgment

This work was supported by National Natural Science Foundation of China (Nos. 62572059, 62376271, U22B2034, 62365014, and 62262043), Beijing Natural Science Foundation (JQ23014), Taishan Scholars Program (No.TSQN202507241), Shandong Provincial Natural Science Foundation for Young Scholars Program (No.ZR2025QC1627) Open Project of Key Laboratory of Computing Power Network and Information Security(No. 2024PY021), Open Project Program of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University (No. VRLAB2025B03), and Qilu University of Technology (Shandong Academy of Sciences) Major Project under Grant (No. 2025ZDZX02).

References

[1] D. Lin, J. Dai, J. Jia *et al.*, "Scribblesup: Scribble-supervised convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3159–3167.

[2] M. Zhou, F. Gao, and R. K.-Y. Tong, "Boundary-enhanced and density-guided contrastive learning for semi/weakly-supervised medical image segmentation," in *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2024, pp. 4088–4091.

[3] L. Xu, W. Ouyang, M. Bennamoun *et al.*, "Leveraging auxiliary tasks with affinity learning for weakly supervised semantic segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 6984–6993.

[4] J. Lee, J. Yi, C. Shin, and S. Yoon, "Bbam: Bounding box attribution map for weakly supervised semantic and instance segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 2643–2652.

[5] J. Chi, Z. Li, H. Wu *et al.*, "Beyond point annotation: A weakly supervised network guided by multi-level labels generated from four-point annotation for thyroid nodule segmentation in ultrasound image," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[6] B. Zhang, J. Xiao, and Y. Zhao, "Dynamic feature regularized loss for weakly supervised semantic segmentation," *Pattern Recognition*, vol. 164, p. 111540, 2025.

[7] S. Chen, C. Wang, R. Xu *et al.*, "SAGE: Spatial-visual adaptive graph exploration for efficient visual place recognition," in *The Fourteenth International Conference on Learning Representations*, 2026.

[8] J. Ahn and S. Kwak, "Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4981–4990.

[9] C. Wang, S. Chen, Y. Song *et al.*, "Focus on local: Finding reliable discriminative regions for visual place recognition," in *Proceedings of the AAAI*, 2025.

[10] X. Han, S. Chen, Z. Fu *et al.*, "Multimodal fusion and vision-language models: A survey for robot vision," *Information Fusion*, p. 103652, 2025.

[11] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," in *First conference on language modeling*, 2024.

[12] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," in *The International Conference on Learning Representations (ICLR)*, 2022.

[13] A. Vaswani, N. Shazeer, N. Parmar *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[14] Y. Liu, Y. Tian, Y. Zhao *et al.*, "Vmamba: Visual state space model," *Advances in neural information processing systems*, vol. 37, pp. 103 031–103 063, 2024.

[15] Z. Xing, T. Ye, Y. Yang *et al.*, "Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation," in *International conference on medical image computing and computer-assisted intervention*, 2024, pp. 578–588.

[16] Z. Wang, J.-Q. Zheng, Y. Zhang *et al.*, "Mamba-unet: Unet-like pure visual mamba for medical image segmentation," *arXiv preprint arXiv:2402.05079*, 2024.

[17] U. Baid, S. Ghodasara, S. Mohan *et al.*, "The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification," *arXiv preprint arXiv:2107.02314*, 2021.

[18] B. Zhou, A. Khosla, A. Lapedriza *et al.*, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[19] Y.-J. Chen, X. Hu, Y. Shi *et al.*, "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*, 2023, pp. 173–182.

[20] P. Khosla, P. Teterwak, C. Wang *et al.*, "Supervised contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 18 661–18 673, 2020.

[21] Z. Qian, K. Li, M. Lai *et al.*, "Transformer based multiple instance learning for weakly supervised histopathology image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2022, pp. 160–170.

[22] R. R. Selvaraju, M. Cogswell, A. Das *et al.*, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[23] H. Wang, Z. Wang, M. Du *et al.*, "Score-cam: Score-weighted visual explanations for convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 24–25.

[24] K. H. Lee, C. Park, J. Oh *et al.*, "Lfi-cam: learning feature importance for better visual explanation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1355–1363.

[25] P.-T. Jiang, C.-B. Zhang, Q. Hou *et al.*, "Layercam: Exploring hierarchical class activation maps for localization," *IEEE Transactions on Image Processing*, vol. 30, pp. 5875–5888, 2021.

[26] J. Chen and E. C. Frey, "Medical image segmentation via unsupervised convolutional neural network," in *Medical Imaging with Deep Learning*, 2020.

[27] M. Futrega, A. Milesi, M. Marcinkiewicz *et al.*, "Optimized u-net for brain tumor segmentation," in *International MICCAI brainlesion workshop*, 2021, pp. 15–29.