



# AME-CAM: Attentive Multiple-Exit CAM for Weakly Supervised Segmentation on MRI Brain Tumor

Yu-Jen Chen<sup>1</sup>(✉), Xinrong Hu<sup>2</sup>, Yiyu Shi<sup>2</sup>, and Tsung-Yi Ho<sup>3</sup>

<sup>1</sup> National Tsing Hua University, Hsinchu, Taiwan  
yujenchen@gapp.nthu.edu.tw

<sup>2</sup> University of Notre Dame, Notre Dame, IN, USA  
{xhu7, yshi4}@nd.edu

<sup>3</sup> The Chinese University of Hong Kong, Hong Kong, China  
tyho@cse.cuhk.edu.hk

**Abstract.** Magnetic resonance imaging (MRI) is commonly used for brain tumor segmentation, which is critical for patient evaluation and treatment planning. To reduce the labor and expertise required for labeling, weakly-supervised semantic segmentation (WSSS) methods with class activation mapping (CAM) have been proposed. However, existing CAM methods suffer from low resolution due to strided convolution and pooling layers, resulting in inaccurate predictions. In this study, we propose a novel CAM method, Attentive Multiple-Exit CAM (AME-CAM), that extracts activation maps from multiple resolutions to hierarchically aggregate and improve prediction accuracy. We evaluate our method on the BraTS 2021 dataset and show that it outperforms state-of-the-art methods.

**Keywords:** Tumor segmentation · Weakly-supervised semantic segmentation

## 1 Introduction

Deep learning techniques have greatly improved medical image segmentation by automatically extracting specific tissue or substance location information, which facilitates accurate disease diagnosis and assessment. However, most deep learning approaches for segmentation require fully or partially labeled training datasets, which can be time-consuming and expensive to annotate. To address this issue, recent research has focused on developing segmentation frameworks that require little or no segmentation labels.

To meet this need, many researchers have devoted their efforts to Weakly-Supervised Semantic Segmentation (WSSS) [21], which utilizes weak supervision, such as image-level classification labels. Recent WSSS methods can be broadly categorized into two types [4]: Class-Activation-Mapping-based (CAM-based) [9, 13, 16, 19, 20, 22], and Multiple-Instance-Learning-based (MIL-based) [15] methods.

---

**Supplementary Information** The online version contains supplementary material available at [https://doi.org/10.1007/978-3-031-43907-0\\_17](https://doi.org/10.1007/978-3-031-43907-0_17).

The literature has not adequately addressed the issue of low-resolution Class-Activation Maps (CAMs), especially for medical images. Some existing methods, such as dilated residual networks [24] and U-Net segmentation architecture [3, 7, 17], have attempted to tackle this issue, but still require many upsampling operations, which the results become blurry. Meanwhile, LayerCAM [9] has proposed a hierarchical solution that extracts activation maps from multiple convolution layers using Grad-CAM [16] and aggregates them with equal weights. Although this approach successfully enhances the resolution of the segmentation mask, it lacks flexibility and may not be optimal.

In this paper, we propose an Attentive Multiple-Exit CAM (AME-CAM) for brain tumor segmentation in magnetic resonance imaging (MRI). Different from recent CAM methods, AME-CAM uses a classification model with multiple-exit training strategy applied to optimize the internal outputs. Activation maps from the outputs of internal classifiers, which have different resolutions, are then aggregated using an attention model. The model learns the pixel-wise weighted sum of the activation maps by a novel contrastive learning method.

Our proposed method has the following contributions:

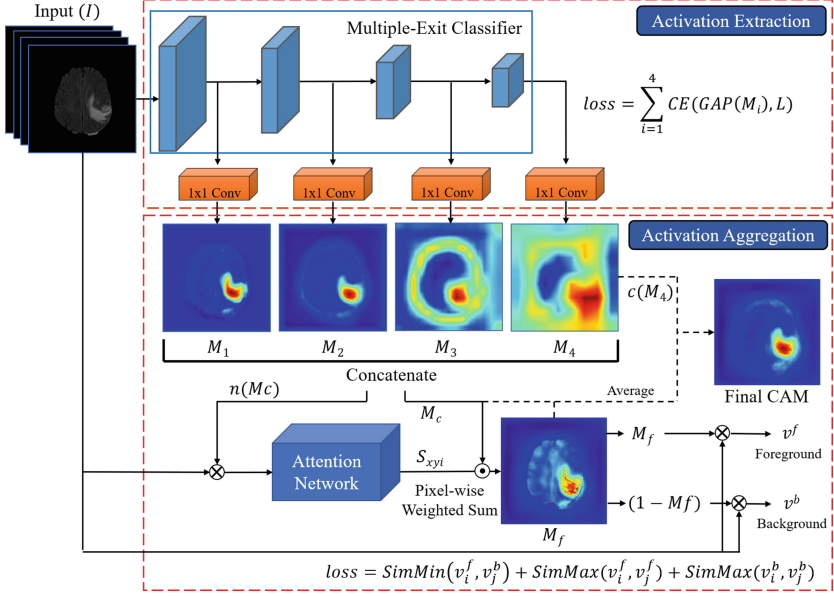
- To tackle the issues in existing CAMs, we propose to use multiple-exit classification networks to accurately capture all the internal activation maps of different resolutions.
- We propose an attentive feature aggregation to learn the pixel-wise weighted sum of the internal activation maps.
- We demonstrate the superiority of AME-CAM over state-of-the-art CAM methods in extracting segmentation results from classification networks on the 2021 Brain Tumor Segmentation Challenge (BraTS 2021) [1, 2, 14].
- For reproducibility, we have released our code at <https://github.com/windstormer/AME-CAM>

Overall, our proposed method can help overcome the challenges of expensive and time-consuming segmentation labeling in medical imaging, and has the potential to improve the accuracy of disease diagnosis and assessment.

## 2 Attentive Multiple-Exit CAM (AME-CAM)

The proposed AME-CAM method consists of two training phases: activation extraction and activation aggregation, as shown in Fig. 1. In the activation extraction phase, we use a binary classification network, e.g., ResNet-18, to obtain the class probability  $y = f(I)$  of the input image  $I$ . To enable multiple-exit training, we add one internal classifier after each residual block, which generates the activation map  $M_i$  of different resolutions. We use a cross-entropy loss to train the multiple-exit classifier, which is defined as

$$loss = \sum_{i=1}^4 CE(GAP(M_i), L) \quad (1)$$



**Fig. 1.** An overview of the proposed AME-CAM method, which contains multiple-exit network based activation extraction phase and attention based activation aggregation phase. The operator  $\odot$  and  $\otimes$  denote the pixel-wise weighted sum and the pixel-wise multiplication, respectively.

where  $GAP(\cdot)$  is the global-average-pooling operation,  $CE(\cdot)$  is the cross-entropy loss, and  $L$  is the image-wise ground-truth label.

In the activation aggregation phase, we create an efficient hierarchical aggregation method to generate the aggregated activation map  $M_f$  by calculating the pixel-wise weighted sum of the activation maps  $M_i$ . We use an attention network  $A(\cdot)$  to estimate the importance of each pixel from each activation map. The attention network takes in the input image  $I$  masked by the activation map and outputs the pixel-wised importance score  $S_{xyi}$  of each activation map. We formulate the operation as follows:

$$S_{xyi} = A([I \otimes n(M_i)]_{i=1}^4) \quad (2)$$

where  $[\cdot]$  is the concatenate operation,  $n(\cdot)$  is the min-max normalization to map the range to  $[0,1]$ , and  $\otimes$  is the pixel-wise multiplication, which is known as image masking. The aggregated activation map  $M_f$  is then obtained by the pixel-wise weighted sum of  $M_i$ , which is  $M_f = \sum_{i=1}^4 (S_{xyi} \otimes M_i)$ .

We train the attention network with unsupervised contrastive learning, which forces the network to disentangle the foreground and the background of the aggregated activation map  $M_f$ . We mask the input image by the aggregated activation map  $M_f$  and its opposite  $(1 - M_f)$  to obtain the foreground feature

and the background feature, respectively. The loss function is defined as follows:

$$loss = SimMin(v_i^f, v_j^b) + SimMax(v_i^f, v_j^f) + SimMax(v_i^b, v_j^b) \quad (3)$$

where  $v_i^f$  and  $v_i^b$  denote the foreground and the background feature of the  $i$ -th sample, respectively.  $SimMin$  and  $SimMax$  are the losses that minimize and maximize the similarity between two features (see C<sup>2</sup>AM [22] for details).

Finally, we average the activation maps  $M_1$  to  $M_4$  and the aggregated map  $M_f$  to obtain the final CAM results for each image. We apply the Dense Conditional Random Field (DenseCRF) [12] algorithm to generate the final segmentation mask. It is worth noting that the proposed method is flexible and can be applied to any classification network architecture.

### 3 Experiments

#### 3.1 Dataset

We evaluate our method on the Brain Tumor Segmentation challenge (BraTS) dataset [1, 2, 14], which contains 2,000 cases, each of which includes four 3D volumes from four different MRI modalities: T1, post-contrast enhanced T1 (T1-CE), T2, and T2 Fluid Attenuated Inversion Recovery (T2-FLAIR), as well as a corresponding segmentation ground-truth mask. The official data split divides these cases by the ratio of 8:1:1 for training, validation, and testing (5,802 positive and 1,073 negative images). In order to evaluate the performance, we use the validation set as our test set and report statistics on it. We preprocess the data by slicing each volume along the z-axis to form a total of 193,905 2D images, following the approach of Kang et al. [10] and Dey and Hong [6]. We use the ground-truth segmentation masks only in the final evaluation, not in the training process.

#### 3.2 Implementation Details and Evaluation Protocol

We implement our method in PyTorch using ResNet-18 as the backbone classifier. We pretrain the classifier using SupCon [11] and then fine-tune it in our experiments. We use the entire training set for both pretraining and fine-tuning. We set the initial learning rate to 1e-4 for both phases, and use the cosine annealing scheduler to decrease it until the minimum learning rate is 5e-6. We set the weight decay in both phases to 1e-5 for model regularization. We use Adam optimizer in the multiple-exit phase and SGD optimizer in the aggregation phase. We train all classifiers until they converge with a test accuracy of over 0.9 for all image modalities. Note that only class labels are available in the training set.

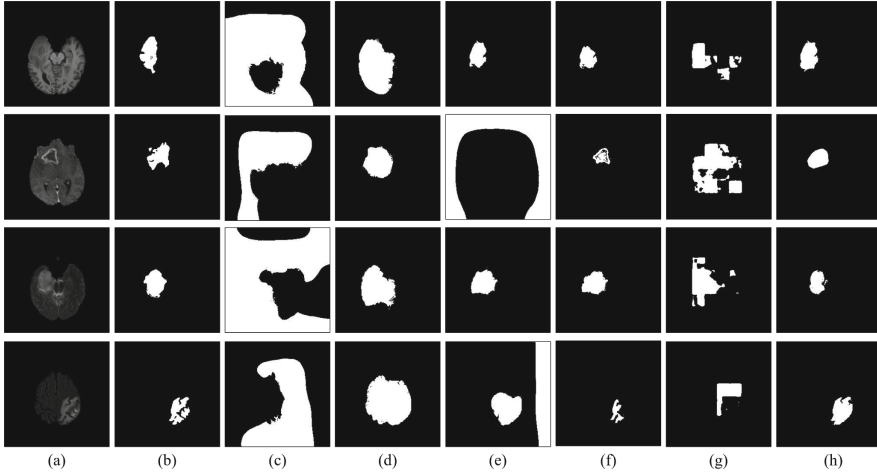
We use the Dice score and Intersection over Union (IoU) to evaluate the quality of the semantic segmentation, following the approach of Xu et al. [23], Tang et al. [18], and Qian et al. [15]. In addition, we report the 95% Hausdorff Distance (HD95) to evaluate the boundary of the prediction mask.

Interested readers can refer to the supplementary material for results on other network architectures.

**Table 1.** Comparison with weakly supervised methods (WSSS), unsupervised method (UL), and fully supervised methods (FSL) on BraTS dataset with T1, T1-CE, T2, and T2-FLAIR MRI images. Results are reported in the form of mean  $\pm$  std. We mark the highest score among WSSS methods with bold text.

BraTS T1				
Type	Method	Dice $\uparrow$	IoU $\uparrow$	HD95 $\downarrow$
WSSS	Grad-CAM (2016)	0.107 $\pm$ 0.090	0.059 $\pm$ 0.055	121.816 $\pm$ 22.963
	ScoreCAM (2020)	0.296 $\pm$ 0.128	0.181 $\pm$ 0.089	60.302 $\pm$ 14.110
	LFI-CAM (2021)	0.568 $\pm$ 0.167	0.414 $\pm$ 0.152	23.939 $\pm$ 25.609
	LayerCAM (2021)	0.571 $\pm$ 0.170	0.419 $\pm$ 0.161	23.335 $\pm$ 27.369
	Swin-MIL (2022)	0.477 $\pm$ 0.170	0.330 $\pm$ 0.147	46.468 $\pm$ 30.408
	AME-CAM (ours)	<b>0.631 <math>\pm</math> 0.119</b>	<b>0.471 <math>\pm</math> 0.119</b>	<b>21.813 <math>\pm</math> 18.219</b>
UL	C&F (2020)	0.200 $\pm$ 0.082	0.113 $\pm$ 0.051	79.187 $\pm$ 14.304
FSL	C&F (2020)	0.572 $\pm$ 0.196	0.426 $\pm$ 0.187	29.027 $\pm$ 20.881
	Opt. U-net (2021)	0.836 $\pm$ 0.062	0.723 $\pm$ 0.090	11.730 $\pm$ 10.345
BraTS T1-CE				
Type	Method	Dice $\uparrow$	IoU $\uparrow$	HD95 $\downarrow$
WSSS	Grad-CAM (2016)	0.127 $\pm$ 0.088	0.071 $\pm$ 0.054	129.890 $\pm$ 27.854
	ScoreCAM (2020)	0.397 $\pm$ 0.189	0.267 $\pm$ 0.163	46.834 $\pm$ 22.093
	LFI-CAM (2021)	0.121 $\pm$ 0.120	0.069 $\pm$ 0.076	136.246 $\pm$ 38.619
	LayerCAM (2021)	0.510 $\pm$ 0.209	0.367 $\pm$ 0.180	29.850 $\pm$ 45.877
	Swin-MIL (2022)	0.460 $\pm$ 0.169	0.314 $\pm$ 0.140	46.996 $\pm$ 22.821
	AME-CAM (ours)	<b>0.695 <math>\pm</math> 0.095</b>	<b>0.540 <math>\pm</math> 0.108</b>	<b>18.129 <math>\pm</math> 12.335</b>
UL	C&F (2020)	0.179 $\pm$ 0.080	0.101 $\pm$ 0.050	77.982 $\pm$ 14.042
FSL	C&F (2020)	0.246 $\pm$ 0.104	0.144 $\pm$ 0.070	130.616 $\pm$ 9.879
	Opt. U-net (2021)	0.845 $\pm$ 0.058	0.736 $\pm$ 0.085	11.593 $\pm$ 11.120
BraTS T2				
Type	Method	Dice $\uparrow$	IoU $\uparrow$	HD95 $\downarrow$
WSSS	Grad-CAM (2016)	0.049 $\pm$ 0.058	0.026 $\pm$ 0.034	141.025 $\pm$ 23.107
	ScoreCAM (2020)	0.530 $\pm$ 0.184	0.382 $\pm$ 0.174	28.611 $\pm$ 11.596
	LFI-CAM (2021)	0.673 $\pm$ 0.173	0.531 $\pm$ 0.186	18.165 $\pm$ 10.475
	LayerCAM (2021)	0.624 $\pm$ 0.178	0.476 $\pm$ 0.173	23.978 $\pm$ 44.323
	Swin-MIL (2022)	0.437 $\pm$ 0.149	0.290 $\pm$ 0.117	38.006 $\pm$ 30.000
	AME-CAM (ours)	<b>0.721 <math>\pm</math> 0.086</b>	<b>0.571 <math>\pm</math> 0.101</b>	<b>14.940 <math>\pm</math> 8.736</b>
UL	C&F (2020)	0.230 $\pm$ 0.089	0.133 $\pm$ 0.058	76.256 $\pm$ 13.192
FSL	C&F (2020)	0.611 $\pm$ 0.221	0.474 $\pm$ 0.217	109.817 $\pm$ 27.735
	Opt. U-net (2021)	0.884 $\pm$ 0.064	0.798 $\pm$ 0.098	8.349 $\pm$ 9.125
BraTS T2-FLAIR				
Type	Method	Dice $\uparrow$	IoU $\uparrow$	HD95 $\downarrow$
WSSS	Grad-CAM (2016)	0.150 $\pm$ 0.077	0.083 $\pm$ 0.050	110.031 $\pm$ 23.307
	ScoreCAM (2020)	0.432 $\pm$ 0.209	0.299 $\pm$ 0.178	39.385 $\pm$ 17.182
	LFI-CAM (2021)	0.161 $\pm$ 0.192	0.102 $\pm$ 0.140	125.749 $\pm$ 45.582
	LayerCAM (2021)	0.652 $\pm$ 0.206	0.515 $\pm$ 0.210	22.055 $\pm$ 33.959
	Swin-MIL (2022)	0.272 $\pm$ 0.115	0.163 $\pm$ 0.079	41.870 $\pm$ 19.231
	AME-CAM (ours)	<b>0.862 <math>\pm</math> 0.088</b>	<b>0.767 <math>\pm</math> 0.122</b>	<b>8.664 <math>\pm</math> 6.440</b>
UL	C&F (2020)	0.306 $\pm$ 0.190	0.199 $\pm$ 0.167	75.651 $\pm$ 14.214
FSL	C&F (2020)	0.578 $\pm$ 0.137	0.419 $\pm$ 0.130	138.138 $\pm$ 14.283
	Opt. U-net (2021)	0.914 $\pm$ 0.058	0.847 $\pm$ 0.093	8.093 $\pm$ 11.879

## 4 Results



**Fig. 2.** Qualitative results of all methods. (a) Input Image. (b) Ground Truth. (c) Grad-CAM [16] (d) ScoreCAM [19]. (e) LFI-CAM [13]. (f) LayerCAM [9]. (g) Swin-MIL [15]. (h) AME-CAM (ours). The image modalities of rows 1-4 are T1, T1-CE, T2, T2-FLAIR, respectively from the BraTS dataset.

### 4.1 Quantitative and Qualitative Comparison with State-of-the-Art

In this section, we compare the segmentation performance of the proposed AME-CAM with five state-of-the-art weakly-supervised segmentation methods, namely Grad-CAM [16], ScoreCAM [19], LFI-CAM [13], LayerCAM [9], and Swin-MIL [15]. We also compare with an unsupervised approach C&F [5], the supervised version of C&F, and the supervised Optimized U-net [8] to show the comparison with non-CAM-based methods. We acknowledge that the results from fully supervised and unsupervised methods are not directly comparable to the weakly supervised CAM methods. Nonetheless, these methods serve as interesting references for the potential performance ceiling and floor of all the CAM methods.

Quantitatively, Grad-CAM and ScoreCAM result in low dice scores, demonstrating that they have difficulty extracting the activation of medical images. LFI-CAM and LayerCAM improve the dice score in all modalities, except LFI-CAM in T1-CE and T2-FLAIR. Finally, the proposed AME-CAM achieves optimal performance in all modalities of the BraTS dataset.

Compared to the unsupervised baseline (UL), C&F is unable to separate the tumor and the surrounding tissue due to low contrast, resulting in low dice scores in all experiments. With pixel-wise labels, the dice of supervised C&F

improves significantly. Without any pixel-wise label, the proposed AME-CAM outperforms supervised C&F in all modalities.

The fully supervised (FSL) Optimized U-net achieves the highest dice score and IoU score in all experiments. However, even under different levels of supervision, there is still a performance gap between the weakly supervised CAM methods and the fully supervised state-of-the-art. This indicates that there is still potential room for WSSS methods to improve in the future.

Qualitatively, Fig. 2 shows the visualization of the CAM and segmentation results from all six CAM-based approaches under four different modalities from the BraTS dataset. Grad-CAM (Fig. 2(c)) results in large false activation region, where the segmentation mask is totally meaningless. ScoreCAM eliminates false activation corresponding to air. LFI-CAM focus on the exact tumor area only in the T1 and T2 MRI (row 1 and 3). Swin-MIL can hardly capture the tumor region of the MRI image, where the activation is noisy. Among all, only LayerCAM and the proposed AME-CAM successfully focus on the exact tumor area, but AME-CAM reduces the under-estimation of the tumor area. This is attributed to the benefit provided by aggregating activation maps from different resolutions.

## 4.2 Ablation Study

**Table 2.** Ablation study for aggregation phase using T1 MRI images from the BraTS dataset. Avg. ME denotes that we directly average four activation maps generated by the multiple-exit phase. The dice score, IoU, and the HD95 are reported in the form of mean  $\pm$  std.

Method	Dice $\uparrow$	IoU $\uparrow$	HD95 $\downarrow$
Avg. ME	$0.617 \pm 0.121$	$0.457 \pm 0.121$	$23.603 \pm 20.572$
Avg. ME+C <sup>2</sup> AM [22]	$0.484 \pm 0.256$	$0.354 \pm 0.207$	$69.242 \pm 121.163$
AME-CAM (ours)	<b><math>0.631 \pm 0.119</math></b>	<b><math>0.471 \pm 0.119</math></b>	<b><math>21.813 \pm 18.219</math></b>

**Effect of Different Aggregation Approaches:** In Table 2, we conducted an ablation study to investigate the impact of using different aggregation approaches after extracting activations from the multiple-exit network. We aim to demonstrate the superiority of the proposed attention-based aggregation approach for segmenting tumor regions in T1 MRI of the BraTS dataset. Note that we only report the results for T1 MRI in the BraTS dataset. Please refer to the supplementary material for the full set of experiments.

As a baseline, we first conducted the average of four activation maps generated by the multiple-level activation extraction (Avg. ME). We then applied C<sup>2</sup>AM [22], a state-of-the-art CAM-based refinement approach, to refine the result of the baseline, which we call “Avg. ME+C<sup>2</sup>AM”. However, we observed that C<sup>2</sup>AM tended to segment the brain region instead of the tumor region due

to the larger contrast between the brain tissue and the air than that between the tumor region and its surrounding tissue. Any incorrect activation of  $C^2AM$  also led to inferior results, resulting in a degradation of the average dice score from 0.617 to 0.484. In contrast, the proposed attention-based approach provided a significant weighting solution that led to optimal performance in all cases.

**Table 3.** Ablation study for using single-exit from  $M_1$ ,  $M_2$ ,  $M_3$  or  $M_4$  of Fig. 1 and the multiple-exit using results from  $M_2$  and  $M_3$  and using all exits (AME-CAM). The experiments are done on the T1-CE MRI of BraTS dataset. The dice score, IoU, and the HD95 are reported in the form of mean  $\pm$  std.

Selected Exit		Dice $\uparrow$	IoU $\uparrow$	HD95 $\downarrow$
Single-exit	$M_1$	$0.144 \pm 0.184$	$0.090 \pm 0.130$	$74.249 \pm 62.669$
	$M_2$	$0.500 \pm 0.231$	$0.363 \pm 0.196$	$43.762 \pm 85.703$
	$M_3$	$0.520 \pm 0.163$	$0.367 \pm 0.141$	$43.749 \pm 54.907$
	$M_4$	$0.154 \pm 0.101$	$0.087 \pm 0.065$	$120.779 \pm 44.548$
Multiple-exit	$M_2 + M_3$	$0.566 \pm 0.207$	$0.421 \pm 0.186$	$27.972 \pm 56.591$
	AME-CAM (ours)	<b><math>0.695 \pm 0.095</math></b>	<b><math>0.540 \pm 0.108</math></b>	<b><math>18.129 \pm 12.335</math></b>

**Effect of Single-Exit and Multiple-Exit:** Table 3 summarizes the performance of using single-exit from  $M_1$ ,  $M_2$ ,  $M_3$ , or  $M_4$  of Fig. 1 and the multiple-exit using results from  $M_2$  and  $M_3$ , and using all exits (AME-CAM) on T1-CE MRI in the BraTS dataset.

The comparisons show that the activation map obtained from the shallow layer  $M_1$  and the deepest layer  $M_4$  result in low dice scores, around 0.15. This is because the network is not deep enough to learn the tumor region in the shallow layer, and the resolution of the activation map obtained from the deepest layer is too low to contain sufficient information to make a clear boundary for the tumor. Results of the internal classifiers from the middle of the network ( $M_2$  and  $M_3$ ) achieve the highest dice score and IoU, both of which are around 0.5.

To evaluate whether using results from all internal classifiers leads to the highest performance, we further apply the proposed method to the two internal classifiers with the highest dice scores, i.e.,  $M_2$  and  $M_3$ , called  $M_2 + M_3$ . Compared with using all internal classifiers ( $M_1$  to  $M_4$ ),  $M_2 + M_3$  results in 18.6% and 22.1% lower dice and IoU, respectively. In conclusion, our AME-CAM still achieves the optimal performance among all the experiments of single-exit and multiple-exit.

Other ablation studies are presented in the supplementary material due to space limitations.

## 5 Conclusion

In this work, we propose a brain tumor segmentation method for MRI images using only class labels, based on an Attentive Multiple-Exit Class Activation



Mapping (AME-CAM). Our approach extracts activation maps from different exits of the network to capture information from multiple resolutions. We then use an attention model to hierarchically aggregate these activation maps, learning pixel-wise weighted sums.

Experimental results on the four modalities of the 2021 BraTS dataset demonstrate the superiority of our approach compared with other CAM-based weakly-supervised segmentation methods. Specifically, AME-CAM achieves the highest dice score for all patients in all datasets and modalities. These results indicate the effectiveness of our proposed approach in accurately segmenting brain tumors from MRI images using only class labels.

## References

1. Bakas, S., et al.: Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific Data* **4**(1), 1–13 (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 preprint [arXiv:1811.02629](https://arxiv.org/abs/1811.02629) (2018)
3. Belharbi, S., Sarraf, A., Pedersoli, M., Ben Ayed, I., McCaffrey, L., Granger, E.: F-cam: Ffull resolution class activation maps via guided parametric upscaling. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3490–3499 (2022)
4. Chan, L., Hosseini, M.S., Plataniotis, K.N.: A comprehensive analysis of weakly-supervised semantic segmentation in different image domains. *Int. J. Comput. Vision* **129**, 361–384 (2021)
5. Chen, J., Frey, E.C.: Medical image segmentation via unsupervised convolutional neural network. arXiv preprint [arXiv:2001.10155](https://arxiv.org/abs/2001.10155) (2020)
6. Dey, R., Hong, Y.: ASC-Net: adversarial-based selective network for unsupervised anomaly segmentation. In: de Bruijne, M., et al. (eds.) *MICCAI 2021. LNCS*, vol. 12905, pp. 236–247. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-87240-3\\_23](https://doi.org/10.1007/978-3-030-87240-3_23)
7. Englebert, A., Cornu, O., De Vleeschouwer, C.: Poly-cam: high resolution class activation map for convolutional neural networks. arXiv preprint [arXiv:2204.13359](https://arxiv.org/abs/2204.13359) (2022)
8. Futrega, M., Milesi, A., Marcinkiewicz, M., Ribalta, P.: Optimized u-net for brain tumor segmentation. arXiv preprint [arXiv:2110.03352](https://arxiv.org/abs/2110.03352) (2021)
9. Jiang, P.T., Zhang, C.B., Hou, Q., Cheng, M.M., Wei, Y.: Layercam: exploring hierarchical class activation maps for localization. *IEEE Trans. Image Process.* **30**, 5875–5888 (2021)
10. Kang, H., Park, H.m., Ahn, Y., Van Messem, A., De Neve, W.: Towards a quantitative analysis of class activation mapping for deep learning-based computer-aided diagnosis. In: *Medical Imaging 2021: Image Perception, Observer Performance, and Technology Assessment*, vol. 11599, p. 115990M. International Society for Optics and Photonics (2021)
11. Khosla, P., et al.: Supervised contrastive learning. arXiv preprint [arXiv:2004.11362](https://arxiv.org/abs/2004.11362) (2020)
12. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected crfs with gaussian edge potentials. In: *Advances in Neural Information Processing Systems 24* (2011)

13. Lee, K.H., Park, C., Oh, J., Kwak, N.: Lfi-cam: learning feature importance for better visual explanation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1355–1363 (2021)
14. Menze, B.H., et al.: The multimodal brain tumor image segmentation benchmark (brats). *IEEE Trans. Med. Imaging* **34**(10), 1993–2024 (2014)
15. Qian, Z., et al.: Transformer based multiple instance learning for weakly supervised histopathology image segmentation. In: Medical Image Computing and Computer Assisted Intervention-MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part II, pp. 160–170. Springer (2022). [https://doi.org/10.1007/978-3-031-16434-7\\_16](https://doi.org/10.1007/978-3-031-16434-7_16)
16. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626 (2017)
17. Tagaris, T., Sdraka, M., Stafylopatis, A.: High-resolution class activation mapping. In: 2019 IEEE International Conference On Image Processing (ICIP), pp. 4514–4518. IEEE (2019)
18. 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: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12907, pp. 262–272. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-87234-2\\_25](https://doi.org/10.1007/978-3-030-87234-2_25)
19. Wang, H., 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, pp. 24–25 (2020)
20. Wang, Y., Zhang, J., Kan, M., Shan, S., Chen, X.: Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12275–12284 (2020)
21. Wolleb, J., Bieder, F., Sandkühler, R., Cattin, P.C.: Diffusion models for medical anomaly detection. In: Medical Image Computing and Computer Assisted Intervention-MICCAI 2022: 25th International Conference, Singapore, 18–22 September 2022, Proceedings, Part VIII, pp. 35–45. Springer (2022). [https://doi.org/10.1007/978-3-031-16452-1\\_4](https://doi.org/10.1007/978-3-031-16452-1_4)
22. Xie, J., Xiang, J., Chen, J., Hou, X., Zhao, X., Shen, L.: C2am: contrastive learning of class-agnostic activation map for weakly supervised object localization and semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 989–998 (2022)
23. Xu, X., et al.: Whole heart and great vessel segmentation in congenital heart disease using deep neural networks and graph matching. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11765, pp. 477–485. Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-32245-8\\_53](https://doi.org/10.1007/978-3-030-32245-8_53)
24. Yu, F., Koltun, V., Funkhouser, T.: Dilated residual networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 472–480 (2017)