

SAM-DRA-UNet: An Enhanced U-Net Framework Integrating Knowledge Distillation and Transfer Learning for Brain Tumor Segmentation

Weihao Huang¹, Chunhong Jiang¹, Yuheng Huang¹, Jiayu Ye², Yuntao Nie¹ and Jiahui Pan^{1✉}

¹ School of Artificial Intelligence, South China Normal University, Foshan, China

² Aberdeen Institute of Data Science and Artificial Intelligence, South China Normal University, Foshan, China
panjh82@qq.com

Abstract. Brain tumor segmentation is challenged by irregular morphology, scarce annotations, and class imbalance in medical imaging. This study proposes SAM-DRA-UNet, an enhanced U-Net framework integrating knowledge distillation and transfer learning. We first develop the DRA-UNet architecture by augmenting U-Net's convolutional blocks with a novel depthwise-pointwise reinforced module and multiple residual simple attention modules, which infer 3D attention maps without parameter expansion while preserving baseline network weights. Furthermore, we employ the SAM model as the teacher network and the DRA-UNet as the student network, transferring knowledge through distillation. Experiments demonstrate that the model achieves mIoU scores of 0.8276 on the TCGA-LGG dataset and 0.8479 on the BraTS21 dataset, significantly outperforming the baseline U-Net and existing state-of-the-art methods. The model also exhibits stable performance across diverse datasets and knowledge distillation temperature settings, validating its generalization capability and providing a reliable solution for brain tumor image segmentation.

Keywords: Brain tumor segmentation, SAM-DRA-UNet, Knowledge distillation, Transfer learning.

1 Introduction

1.1 A Subsection Sample

Brain tumors are neoplasms that develop within the cranial cavity, which may arise either from metastatic invasion of extracranial organs or tissues into the intracranial space or directly originate from brain tissue, nerves, meninges, cerebral appendages, or blood vessels. Brain tumor segmentation is a pivotal task in medical image processing, aiming to accurately delineate and identify tumor lesion regions from medical imaging data. This process typically relies on multimodal magnetic resonance imaging (MRI) sequences, such as T1-weighted (T1), contrast-enhanced T1-weighted

(T1-Gd), T2-weighted (T2), and T2 Fluid Attenuated Inversion Recovery (T2-FLAIR), which provide complementary clinical information for tumor analysis [1].

Despite continuous advancements in brain tumor imaging technologies, segmentation tasks remain challenged by morphological ambiguity, annotation bias from scarce labeled data, and class imbalance. Tumor infiltration into surrounding healthy tissues creates blurred boundaries between pathological and normal regions. Gliomas, originating from widely distributed glial cells, may develop anywhere in the brain with substantial morphological variations across disease stages and individual patients. Manual annotations suffer from operator-dependent subjectivity and high inter-observer variability, compounded by the limited availability of annotated datasets due to the exorbitant costs of medical image labeling. Furthermore, the imbalanced voxel distribution across tumor subregions demands architectures capable of precisely resolving intricate imaging boundaries.

To address these challenges, we propose SAM-DRA-UNet, an enhanced brain tumor segmentation model integrating knowledge distillation and transfer learning. The model combines the powerful generalization capability of the foundational SAM [2] model with the efficient feature extraction advantages of a lightweight improved U-Net [3] to achieve precise tumor boundary delineation. Specifically, we first develop an enhanced U-Net called DRA-UNet (Depthwise-Residual-Attention UNet), which incorporates a depthwise-pointwise reinforced module (DPRM) and multiple residual simple attention modules (Res-SimAM). In the encoder, DPRM employs cascaded depthwise and pointwise convolutions to reduce computational complexity while preserving shallow-layer details via residual connections. Res-SimAM, deployed in both encoder and decoder stages, dynamically calibrates multi-scale features through energy function-derived spatial-channel attention weights without introducing additional parameters. Furthermore, we establish a cross-model knowledge transfer framework. This framework leverages the large-scale pre-trained SAM model as the teacher network and implements a hybrid soft-hard label distillation strategy integrating Kullback-Leibler divergence and cross-entropy, effectively transferring semantic boundary perception capabilities to the lightweight DRA-UNet student network. Trained on publicly available TCGA-LGG [4] and BraTS21 [5] datasets, the model demonstrates exceptional performance under limited annotation conditions, validating its robustness in handling complex tumor morphology and low-contrast boundaries for real-world medical imaging applications.

Overall, our main contributions are summarized as follows:

- We facilitate efficient knowledge transfer from the large-scale pre-trained general segmentation model SAM to the lightweight DRA-UNet, achieving a significant reduction in computational demands while maintaining high-precision tumor boundary segmentation.
- We propose two novel modules: Depthwise-Pointwise Reinforced Module and Residual Simple Attention Module. These modules enhance the cross-scale modeling capability of U-Net through lightweight convolutions and adaptive feature calibration, without introducing additional parameters.

- We develop a collaborative soft-hard label distillation framework integrating Kullback-Leibler divergence and cross-entropy loss, which enables efficient cross-domain boundary-aware knowledge transfer tailored for medical imaging’s complex morphological structures and low-contrast boundaries.
- Extensive experiments demonstrate the effectiveness and generalization ability of the proposed SAM-UNet and show new state-of-the-art results on two challenging datasets, i.e., TCGA-LGG and BraTS21.

2 Related Work

Early brain tumor segmentation relied on manual delineation of MRI scans by experts, which suffered from inter-operator variability. While automated methods later emerged, most algorithms were validated on heterogeneous private datasets with varying modalities, tumor types, and disease stages, hindering objective performance comparisons.

In recent years, the BraTS challenge initiated by MICCAI has significantly advanced brain tumor segmentation technologies. Early techniques, such as threshold-based segmentation, region-growing, edge detection, and level-set methods—which required substantial prior knowledge—have gradually transitioned to optimization-based approaches like graph cuts and Markov random fields, though their generalization capabilities remained limited. With advancements in pattern recognition, methods such as Support Vector Machines (SVM) [6], Random Forests (RF) [7], and K-means clustering were employed for brain tumor segmentation. U-Net and its variants (e.g., 3D U-Net [8], Attention U-Net [9]) emerged as mainstream models, effectively capturing multi-scale features to improve segmentation accuracy. Additionally, nnU-Net [10] demonstrated outstanding performance in international competitions like BraTS through automated hyperparameter optimization. More recently, Transformer architectures [11] and self-supervised learning have been introduced into the field, further enhancing model generalization. While these methods improved segmentation outcomes by leveraging features such as texture, shape, and grayscale histograms, their reliance on handcrafted feature engineering limited adaptability to heterogeneous patient imaging data.

Current approaches exhibit trade-offs between model efficiency and segmentation precision. R. Zhou et al. [12] proposed a cascaded CNN-Mamba model but overlooked validation on small tumor subregions and cross-dataset generalization. L. Liu et al. [13] developed a lightweight U-Net with 3D depthwise separable convolutions and dilated dense residual blocks, yet its expanded receptive field compromises local details and sensitivity to microscopic tumors. X. Wu et al. [14] enhanced feature propagation via dense U-Net-DenseNet connections and hybrid loss functions to address class imbalance, though dense connectivity increases computational costs with suboptimal boundary precision. W.A. Yang et al. [15] reduced computational overhead using Fermi normalization and fDDFT-based global modules but sacrificed local detail retention through downsampling. Z. Zhu et al. [16] introduced SDV-TUNet with sparse dynamic encoders for multi-level feature fusion,

yet its dynamic sparse attention increases complexity with unverified low-contrast performance. X. Siyi et al. [17] improved U-Net via grouped convolutions and attention mechanisms, constrained by fixed grouping strategies. P. Li et al. [18] fused multi-modal MRI features through multi-scale residuals and channel attention but failed to adapt to inter-modal heterogeneity or model dynamic dependencies.

To address the challenges of poor generalization, insufficient detail capture, and high computational complexity in brain tumor segmentation, this study innovatively integrates the powerful generalization capability of SAM with the lightweight DRA-UNet through a teacher-student training framework. The method combines a depthwise-pointwise reinforced module and multiple residual simple attention modules with the UNet architecture, then transfers SAM’s global segmentation knowledge via knowledge distillation while leveraging SAM’s prompt encoder to enhance the identification of micro-tumors and low-contrast regions. This framework achieves a balance between segmentation accuracy and computational efficiency. By retaining SAM’s cross-domain adaptability and inheriting DRA-UNet’s parameter-efficient design, it ensures accurate tumor boundary delineation in resource-constrained clinical environments.

3 Method

3.1 Overview

The overview of our model is shown in Fig. 1. It consists of a teacher-student architecture designed for efficient and accurate mask generation. The teacher model is based on SAM, which includes an image encoder to process the input image, a prompt encoder to incorporate user-defined prompts, and a mask decoder that outputs segmentation masks through a multi-stage refinement process.

The student model, DRA-UNet, receives the same image and prompt inputs. It leverages a depthwise-pointwise reinforced module (DPRM) to capture richer contextual features and residual simple attention modules (Res-SimAM) to enhance focus on salient regions. Through knowledge distillation, the student model learns from the teacher’s predictions by aligning both soft labels and hard labels.

During training, the model computes the distillation loss by combining Kullback-Leibler divergence for soft labels and cross-entropy for hard labels. The total loss, formed as a weighted sum, guides the optimization of the student model via backpropagation. Ultimately, the trained DRA-UNet can independently generate accurate segmentation masks given an input image and prompt.

3.2 Teacher Model

Segment Anything Model (SAM) is a universal, zero-shot segmentation framework requiring no task-specific training. It processes diverse image types (e.g., medical scans) via end-to-end architecture, leveraging large-scale pre-training and promptable design to localize microscopic tumors and low-contrast regions with high precision.

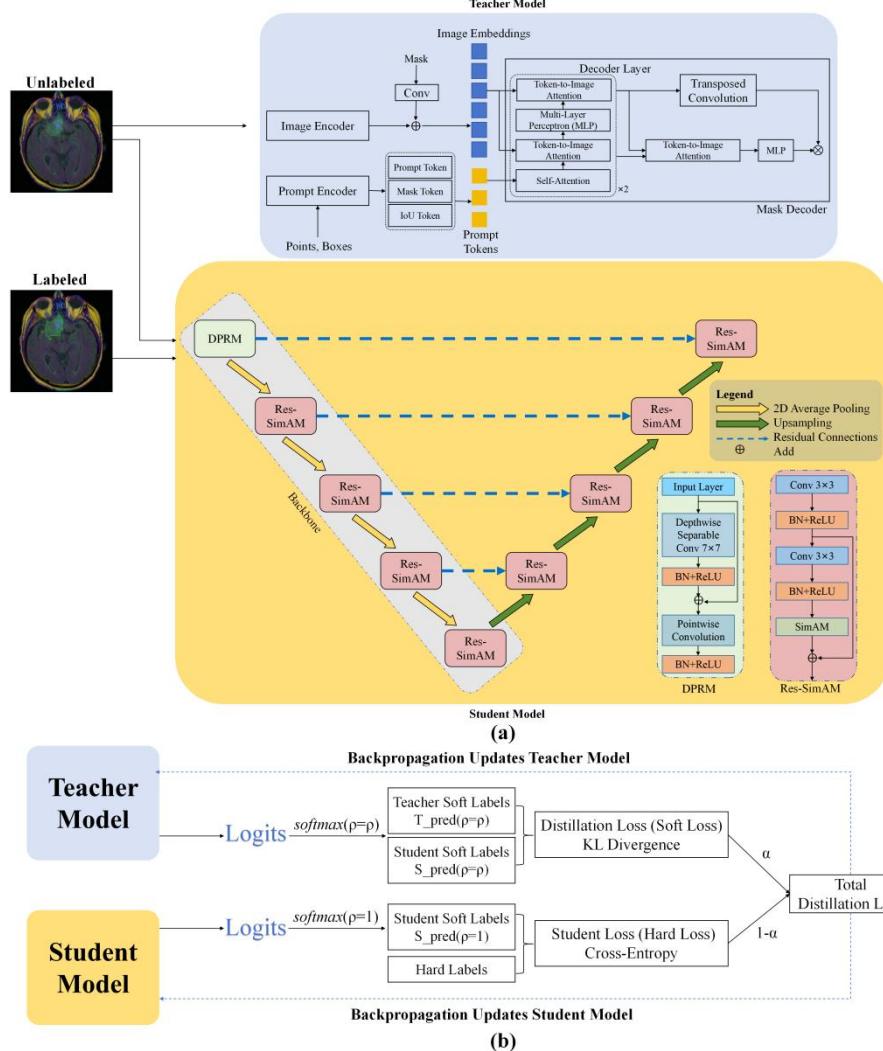


Fig. 1. Overview of our method. (a) SAM is used as the teacher model, with its image encoder encoding images, prompt encoder handling input prompts, and mask decoder generating masks via multi-step operations. DRA-UNet serves as the student model. The large acceptance area of DPRM allows the model to extract denser feature information, while Res-SimAM allows the model to focus on key features. The two models perform knowledge distillation, where the student model learns from the teacher model using soft and hard labels, and updates its parameters through backpropagation for training. (b) The loss of SAM's soft label and DRA-UNet's soft and hard label is calculated by Kullback-Leibler (KL) dispersion and cross entropy (CE), and the weighted sum is used to obtain the distillation total loss, which is used to update the student model parameters.

Image Encoder. The image encoder processes input images using a MAE-pre-trained ViT [19] as the backbone network. It accepts unlabeled input images, processes each image once, and extracts deep features to generate high-dimensional image embeddings, mapping the input image into a feature space. For an input image of size (C, H, W) , it is first divided into non-overlapping patches of size $(C, 1024, 1024)$. These patches are then linearly transformed into tokens of size $(256, 64, 64)$:

$$z_0 = [x_1 E; x_2 E; \dots; x_N E] + E_{pos} \quad (1)$$

Where, x_i represents the i -th image patch, E is the linear transformation matrix converting patches to tokens, and E_{pos} denotes positional encoding. These embeddings are fed into the decoder, where they fuse with prompt information to guide the segmentation task.

Prompt Encoder. The prompt encoder encodes user-provided inputs (e.g., points, bounding boxes) into task-specific embedding vectors. Prompt tokens convert input prompts into processable vector representations, while mask tokens specialize in predicting segmentation masks, and Intersection-over-Union (IoU) tokens evaluate prediction accuracy. These tokens are transmitted to the decoder layers to guide the model’s attention toward target regions. For point inputs, a fixed embedding vector P_{point} is assigned. For bounding box inputs, the coordinate vector is directly incorporated into Transformer computations as follows:

$$P_{box} = [x_{min}, y_{min}, x_{max}, y_{max}] \quad (2)$$

The generated prompt features are transmitted to the decoder layers to guide the model’s attention to specific regions.

Mask Decoder. The Mask Decoder employs a Transformer architecture with dual decoding layers to process encoder-derived information and produce the final segmentation mask. Cross-attention mechanisms mediate bidirectional interactions between semantic tokens and image features, where self-attention refines contextual fusion through inter-token relationships, while token-to-image attention dynamically aligns spatial features to prioritize target regions. This hierarchical attention framework enables adaptive focus on diagnostically relevant areas through feature-token correlation mapping.

A multilayer perceptron performs linear transformations for feature conversion and format adaptation. The token-to-image attention directs the decoder to focus on diagnostically critical regions through cross-attention-like computations, where prompt features dynamically steer image embeddings toward target areas.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (3)$$

$$O = AV \quad (4)$$

Q is derived from prompt features, while K and V originate from image embeddings. This mechanism enables the model to focus exclusively on prompted regions, improving segmentation accuracy. Deconvolution enhances mask resolution, and tokens are processed as queries via attention with image embeddings. The mask token is extracted, passed through MLP, and multiplied with embeddings to generate the final mask. An additional IoU token within the token sequence is processed through MLP to predict confidence scores. Formulas for deconvolution and MLP:

$$O(i, j) = \sum_0^{M-1} \sum_0^{N-1} I(i - m, j - n) \cdot K(m, n) \quad (5)$$

$$y = \sigma(W_2 \cdot \text{ReLU}(W_1 x + b_1) + b_2) \quad (6)$$

In the deconvolution operation, $I(i, j)$ represents the input features, $K(m, n)$ denotes the convolution kernel, and $O(i, j)$ corresponds to the output high-resolution features. For MLP processing, W_1 and W_2 are weight matrices, b_1 and b_2 are bias terms, and $\sigma(\cdot)$ is the activation function. The final binarized mask for target region segmentation is generated as:

$$\hat{M} = \sigma(W_M x + b_M) \quad (7)$$

Where W is the weight matrix, x denotes the decoder output features, and $\sigma(\cdot)$ refers to the Sigmoid function that maps results to the range [0,1].

3.3 Student Model

The student model adopts DRA-UNet, an enhanced U-Net, with five hierarchical stages: a backbone encoder followed by upsampling decoders. Identical-dimensional convolutional tensors per layer and 2×2 max-pooling are utilized. The encoder integrates custom depthwise-pointwise reinforced module (DPRM) and residual simple attention modules (Res-SimAM), while the decoder employs Res-SimAM-enhanced dense skip connections to prioritize critical texture features.

Depthwise-Pointwise Reinforced Module. The depthwise-pointwise reinforced module (DPRM) builds upon the principles of depthwise separable convolution, a technique widely adopted in efficient neural network designs including Xception [20] and MobileNet [21]. Within the encoding pathway, conventional convolution operations are substituted with our enhanced DPRM implementation.

The module first processes input features through a depthwise convolutional layer employing an expanded 7×7 kernel size with 64 output channels, maintaining spatial dimensions through symmetric padding of 3 pixels. Notably, this operation is performed without bias terms to reduce parameter redundancy. A residual connection then combines the processed features with the original input, promoting information flow. Subsequent processing involves batch normalization and ReLU activation, followed by a parameter-efficient pointwise convolution that similarly omits bias terms.

Mathematically, the module's operations can be expressed as:

$$f_1 = x + D(\sigma\{B(x)\}) \quad (8)$$

$$f_2 = P(\sigma\{B(f_1)\}) \quad (9)$$

Where, f_1 represents the intermediate feature map, $\sigma\{\cdot\}$ denotes the ReLU nonlinearity, $B(\cdot)$ signifies batch normalization, x corresponds to the input tensor, $D(\cdot)$ indicates depthwise convolution, and $P(\cdot)$ refers to pointwise convolution. This design enables the model to thoroughly extract image features at the initial stage of the encoding phase, thereby providing rich hierarchical information for subsequent residual simple attention modules. Compared to conventional 3×3 convolutions, the proposed method exhibits a larger receptive field, which enhances dense feature capture while better balancing global context awareness—a critical advantage for processing brain tumor imaging data. Additionally, the residual pathway facilitates gradient propagation from shallow to deep layers, effectively mitigating overfitting risks by preserving low-level texture details during feature abstraction.

Residual Simple Attention Module. The residual simple attention module (Res-SimAM) integrates the SimAM attention mechanism — a neuroscience-inspired, parameter-free 3D attention method (Qin et al. [22]). Initially proven in speaker verification, SimAM adaptively weights spatial-channel features to suppress noise and highlight discriminative patterns, making it particularly effective for medical imaging where subtle tissue variations dictate diagnostic accuracy.

The Res-SimAM uses two stacked 3×3 convolutional blocks (stride=1, padding=1), each with BN and ReLU. The second block uniquely appends a parameter-free SimAM attention mechanism. Pre- and post-attention features are element-wise summed as the output, enhancing semantic feature fusion for segmentation tasks. The Moudule is defined as:

$$g_1 = \sigma(B\{Conv(x)\}) \quad (10)$$

$$g_2 = S\{\sigma(B\{Conv(f_1)\})\} \quad (11)$$

$$g_3 = g_1 + g_2 \quad (12)$$

Where, g_3 denotes the output of the Res-SimAM, $\sigma(\cdot)$ denotes the ReLU nonlinearity, $B\{\cdot\}$ signifies batch normalization, $Conv(\cdot)$ corresponds to the 3×3 convolutional layer, and $S\{\cdot\}$ signifies the SimAM attention mechanism. During encoding, standard convolution blocks in downsampling stages are replaced with residual simple attention modules. This substitution enables adaptive filtering and prioritization of complex features, directing the model's focus toward discriminative patterns while mitigating local optima entrapment. In the decoder, upsampling blocks retain the original UNet architecture with dense skip connections, ensuring consistency with the UNet framework's reconstruction process.

Residual Connection Mechanism. The residual connection mechanism introduces skip connections within convolutional networks by directly adding the input features

to the output of deeper layers, forming a shortcut path. This design alleviates the gradient vanishing problem in deep architectures while preserving shallow-layer feature information, thereby enhancing the model's representational capacity. Let x denote the input features and $F(x)$ represent the features extracted through two convolutional layers. The output of the residual connection can be expressed as:

$$y = x + \zeta(x) \quad (13)$$

Where, $\zeta(x)$ represents nonlinear features extracted by convolutional layers, while x is the original input transmitted via skip connections. This design allows simultaneous capture of deep semantic patterns and retention of low-level details, enriching feature diversity. The residual addition operation, parameter-free and computationally efficient, directly propagates shallow features to deeper layers, optimizing pixel-level classification in segmentation tasks.

3.4 Knowledge Distillation and Transfer Learning

Following the knowledge distillation framework, we utilize the pre-trained, highly robust large model SAM as the teacher model, and a trainable lightweight small model DRA-UNet as the student model, performing transfer learning through distilled feature extraction. The student model combines soft labels and hard labels during distillation to mitigate error propagation. Training proceeds via backpropagation: gradients computed from the loss functions update student model parameters. The Kullback-Leibler divergence loss, cross-entropy loss, and total distillation loss are formulated as:

$$q_i(z_i; \rho) = \frac{\exp(z_i/\rho)}{\sum_{j=0}^n \exp(z_j/\rho)} \quad (14)$$

$$\text{KL}(q(z^T; \rho) || q(z^S; \rho)) = \sum_{i=1}^n q(z_i^T; \rho) \log(\frac{q(z_i^T; \rho)}{q(z_i^S; \rho)}) \quad (15)$$

$$\text{CE}(z^{hard}, z^S) = -\sum_x^{hard}(x) \log(z^S(x)) \quad (16)$$

$$L_{KD} = \alpha \text{KL}(q(z^T; \rho) || q(z^S; \rho)) + (1 - \alpha) \text{CE}(z^{hard}, z^S) \quad (17)$$

Where z^S is the logits-fusion output of the student model, z^{hard} is the hard-label (true label), z^T is the logits-fusion output of the teacher model, ρ is the temperature coefficient, and α is the balance coefficient.

After each distillation iteration, the distillation loss and student loss are weighted and summed, then fed back to the DRA-UNet student model for automatic MRI brain tumor segmentation. Through knowledge transfer, the learned knowledge from source tasks in public datasets can be migrated to target tasks in real-world application scenarios, improving model performance in practical environments. By calculating the total distillation loss and continuously back-propagating it between the teacher and student models, the student model is expected to enhance its generalization

capability from SAM's rich knowledge, achieving robust performance across diverse real-world scenarios.

4 Experiments

4.1 Datasets and Evaluation Metrics

- **TCGA-LGG dataset:** It is dedicated to low-grade glioma (LGG) segmentation, comprising multimodal MRI volumes (T1, T1-Gd, T2, T2-FLAIR) with annotations for tumor subregions. These annotations differentiate enhancing tumor (ET), whole tumor (WT), and tumor core (TC) using color-coded labels (black: background, light gray: non-enhancing tumor, light gold: edema, blue: ET). The training set includes 387 MRI slices from 65 patients. Raw data and lesion masks are stored in *.tif* and *.mask.tif* formats, respectively, and visualized via 3D Slicer software.
- **BraTS21 dataset:** It contains 8,160 multimodal MRI scans (T1, T1-Gd, T2, T2-FLAIR) from 2,040 patients, annotated with four categories: background (black), enhancing tumor (blue), edema (light gold), and necrotic tumor core (green). The dataset is partitioned into 1,251 training cases with public labels, 219 validation cases, and 570 test cases. Both images and masks are stored in *.nii* format and analyzed using ITK-SNAP.
- Evaluation Metrics: The **mIoU** measures the average overlap between predicted and ground-truth regions, providing a robust assessment of segmentation accuracy across all tumor classes. The **F1-score** balances precision and recall, emphasizing the model's ability to correctly identify tumor pixels while minimizing false positives and negatives, which is critical for clinical applications. Finally, **OA** quantifies the global pixel-wise classification accuracy but may be biased if the background dominates the image. Together, these metrics comprehensively assess the model's segmentation quality from regional, target-sensitive, and holistic perspectives.

4.2 Implementation Details

Experiments were conducted on an RTX 3090 GPU with knowledge distillation temperature $\tau = 3$. The model was trained using SGD optimizer combined with a ReduceLROnPlateau scheduler for dynamic learning rate adjustment, and optimized via Dice loss to maximize segmentation overlap. Input images were preprocessed to 256×256 resolution, standardized (channel-wise mean=0, std=1), and median-filtered. Data augmentation included random horizontal or vertical flips and brightness variations to improve robustness.

4.3 Experimental Results and Analysis

A comparative analysis was conducted between the brain tumor segmentation results generated by the system and the manual segmentation results (denoted as GT) provided by physicians. Two patient cases were selected as examples in Fig. 2. Both datasets achieved an overall diagnostic accuracy of 100% in distinguishing positive or negative cases, with the model attaining an mIoU of 0.8276 on TCGA-LGG and a higher mIoU of 0.8479 on BraTS21.

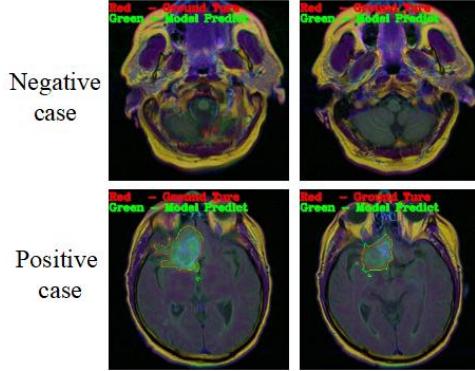


Fig. 2. The segmentation result of our model. Images without red regions indicate negative diagnostic outcomes, while those with red regions represent positive diagnoses. The red regions correspond to lesion areas manually annotated by physicians, and the green regions represent lesion areas segmented by the proposed model.

Under standardized experimental protocols, we conducted benchmark evaluations of multiple brain tumor segmentation models. As detailed in Table 1, on the TCGA-LGG dataset, our model achieves an mIoU of 0.8276, an F1-score of 0.9158, and an OA of 0.9393, surpassing the baseline U-Net by 0.2164, 0.2411, and 0.2402 in these metrics respectively. On the BraTS21 dataset, SAM-DRA-UNet attains an mIoU of 0.8479, an F1-score of 0.9201, and an OA of 0.9407, outperforming the baseline U-Net by 0.2606, 0.2685, and 0.2565 across the same metrics. These results conclusively demonstrate that our model significantly outperforms both the baseline U-Net and other contemporary state-of-the-art methods in mIoU, F1-score, and OA.

Fig. 3 and Fig. 4 illustrate segmentation performance across methods on representative cases. The baseline U-Net generates indistinct tumor localization with inadequate internal detail resolution. MANet enhances boundary clarity but fails to characterize internal structures. LinkNet and YOLOv5s-ASPP surpass these in tumor boundary delineation and subregion partitioning, capturing partial details in enhancing tumor (ET) and edema (ED) areas. Models integrating attention mechanisms or enhanced architectures, including MWG-UNet++, HCA-former, and PyQDCNN, leverage global features to achieve precise non-enhancing core Net and necrosis NCR segmentation. However, minor discrepancies persist in certain textural details. Our model exhibits superior congruence with ground truth in holistic contour matching, internal subregion partitioning, and textural preservation, particularly

excelling in cases with complex morphology, heterogeneous textures, and ambiguous boundaries.

Table 1. Different brain tumor segmentation models were compared on the TCGA-LGG dataset and the BraTS21 dataset.

Dataset	Method	Evaluation Metrics		
		mIoU	F1-score	OA
TCGA-LGG	U-Net [3]	0.6112	0.6747	0.6991
	MANet [23]	0.8094	0.8543	0.8947
	LinkNet [24]	0.7343	0.8212	0.8584
	YOLOv5s-ASPP [25]	0.7264	0.8351	0.8681
	MWG-UNet++ [26]	0.8122	0.9023	0.9217
	HCA-former [27]	0.7976	0.8621	0.9072
	PyQDCNN [28]	0.8046	0.8858	0.9235
	SAM-DRA-UNet (Ours)	0.8276	0.9158	0.9393
BraTS21	U-Net [3]	0.5873	0.6516	0.6842
	MANet [23]	0.7821	0.8575	0.8819
	LinkNet [24]	0.7198	0.8034	0.8115
	YOLOv5s-ASPP [25]	0.7582	0.8427	0.8424
	MWG-UNet++ [26]	0.8205	0.9089	0.9162
	HCA-former [27]	0.8175	0.9086	0.9043
	PyQDCNN [28]	0.8039	0.8721	0.8956
	SAM-DRA-UNet (Ours)	0.8479	0.9201	0.9407

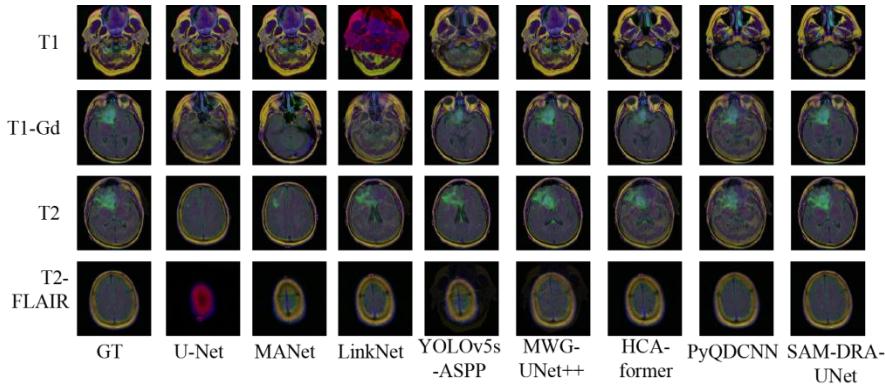


Fig. 3. Segmentation results of different methods on the TCGA_CS_4941 case.

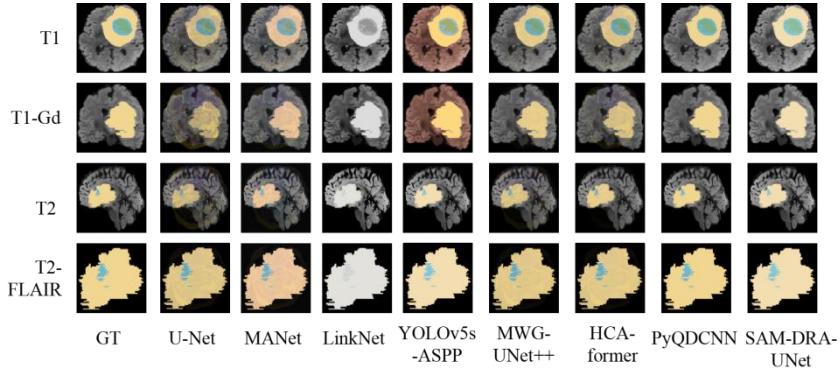


Fig. 4. Segmentation results of different methods on the BraTS21_00002 case.

4.4 Ablation Experiment

To evaluate the effectiveness of our module design, we conducted ablation experiments to analyze in depth the effects of different strategies on model performance. These experiments were carried out in a strict and uniform experimental setting to ensure the fairness and comparability of the results.

Table 2. Results of student model ablation experiment. The last model is student model.

Model	TCGA-LGG dataset			BraTS21 dataset		
	mIoU	F1-score	OA	mIoU	F1-score	OA
U-Net [3]	0.6112	0.6747	0.6991	0.5873	0.6516	0.6842
U-Net + Res-SimAM	0.6634	0.7772	0.7983	0.6421	0.7638	0.7835
U-Net + DPRM	0.6489	0.7426	0.7857	0.6378	0.7583	0.7712
DRA-UNet	0.7224	0.8332	0.8498	0.7035	0.8217	0.8324

Table 3. Ablation experiment results of different knowledge distillation strategies.

Model	TCGA-LGG dataset			BraTS21 dataset		
	mIoU	F1-score	OA	mIoU	F1-score	OA
DRA-UNet	0.7224	0.8332	0.8498	0.7035	0.8217	0.8324
T	0.7386	0.8384	0.8621	0.7298	0.8352	0.8453
T→S (remove soft targets)	0.7149	0.8012	0.8243	0.6967	0.7989	0.8186
T→S (remove hard targets)	0.7721	0.8743	0.8824	0.7586	0.8632	0.8665
T→S (different τ)	0.8235	0.9026	0.9364	0.8342	0.9087	0.9123
T→S (Ours)	0.8276	0.9158	0.9393	0.8479	0.9201	0.9247

Table 2 demonstrates that integrating DPRM and Res-SimAM significantly enhances model performance. On TCGA-LGG, DPRM improves mIoU, F1-score, and OA by 0.0377, 0.0679, and 0.0866; Res-SimAM elevates them by 0.0522, 0.1025, and 0.0992; their combination achieves gains of 0.1112, 0.1585, and 0.1507. Similarly, on BraTS21, DPRM increases these metrics by 0.0505, 0.1067, and 0.0870; Res-SimAM raises them by 0.0548, 0.1122, and 0.0993; combined integration boosts performance to 0.1162, 0.1401, and 0.1482. DPRM mitigates U-Net’s overfitting via expanded receptive fields for contextual integration, while Res-SimAM filters irrelevant features through attention-driven purification. Using DRA-UNet as a student model can better learn the powerful segmentation ability of teacher model on the basis of better effect.

After validating the contributions of DPRM and Res-SimAM to U-Net, we conducted ablation experiments using the DRA-UNet with student loss (cross entropy) as the baseline model to evaluate the impact of different knowledge distillation strategies on segmentation performance, with results shown in Table 3. The $T \rightarrow S$ configuration represents the complete standard knowledge distillation model; $T \rightarrow S$ (remove hard labels) uses only the distillation loss to observe the effect of soft targets; $T \rightarrow S$ (remove soft targets) trains the student model solely with ground truth labels to analyze the contribution of distillation objectives; T denotes using only the teacher model SAM to validate its upper-bound performance; $T \rightarrow S$ (different τ) adjusts the temperature parameter to a value distinct from the initial setting, investigating its influence on distillation efficacy.

The ablation results demonstrate that our complete distillation model ($T \rightarrow S$) achieves superior performance, with mIoU scores of 0.8276 on TCGA-LGG and 0.8479 on BraTS21, significantly surpassing other configurations. This indicates the full distillation process effectively enhances segmentation quality, as evidenced by the consistent improvements in all metrics.

The standalone teacher model T demonstrates intermediate performance, with mIoU scores of 0.7386 on TCGA-LGG and 0.7298 on BraTS21. This positions it between the baseline DRA-UNet’s 0.7224 on TCGA-LGG and 0.7035 on BraTS21, confirming its value as feature guidance while highlighting the student model’s enhanced learning capability. Notably, removing distillation components leads to clear degradation: omitting soft targets reduces mIoU to 0.7149 on TCGA-LGG and 0.6967 on BraTS21, while excluding hard labels yields 0.7721 on TCGA-LGG and 0.7586 on BraTS21 - both configurations proving inferior to our full approach.

The $T \rightarrow S$ (different τ) variant achieves competitive results, reaching 0.8235 mIoU on TCGA-LGG and 0.8342 mIoU on BraTS21. However, the marginal performance gap of less than 2% in mIoU compared to our final model validates the robustness of our default temperature setting. These results collectively confirm the model’s high accuracy and generalization ability for brain MRI segmentation, reliably delivering precise results for clinical analysis across both datasets.

5 Conclusion

In this paper, we propose SAM-DRA-UNet, a novel brain tumor segmentation framework. By integrating the depthwise-pointwise reinforced module and residual simple attention module, we construct DRA-UNet to achieve parameter-efficient feature learning. Additionally, a hybrid knowledge distillation loss combining Kullback-Leibler divergence and cross-entropy is designed to transfer knowledge from the teacher model SAM to the student model DRA-UNet. Extensive experimental results demonstrate that our model outperforms previous state-of-the-art methods on both TCGA-LGG and BraTS21 datasets, while exhibiting strong generalization capability, thereby providing a reliable solution for brain tumor image analysis.

References

1. Xing, Z., Yu, L., Wan, L., Han, T., Zhu L.: Nestedformer: Nested modality-aware transformer for brain tumor segmentation pp. 140-150 (2022)
2. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything pp. 4015-4026 (2023)
3. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation pp. 234-241 (2015)
4. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C.: Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data* **4**(1), 1-13 (2017)
5. Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., Cola, E., Farahani, K., Kalpathy-Cramer, J., Kitamura, F.C., Pati, S., et al.: The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2017.02314* (2021)
6. Hearst, M.A., Dumais, S.T., Osuna, E., Platt, J., Scholkopf, B.: Support vector machines. *IEEE Intelligent Systems and their applications* **13**(4), 18-28 (1998)
7. Rigatti, S.J.: Random forest. *Journal of Insurance Medicine* **47**(1), 31-39 (2017)
8. Çiçek Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation pp. 424-432 (2016)
9. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al.: Attention u-net; Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999* (2018)
10. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnU-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods* **18**(2), 203-211 (2021)
11. Lyu, H., Sha, N., Qin, S., Yan, M., Xie, Y., Wang, R.: Advances in neural information processing systems. *Advances in neural information processing systems* **32** (2019)
12. Zhou, R., Wang, J., Xia, G., Xing, J., Shen, H., Shen, X.: Cascade residual multiscale convolution and mamba-structured unet for advanced brain tumor image segmentation. *Entropy* **26**(5), 385 (2024)
13. Liu, L., Xia, K.: Btis-net: Efficient 3d u-net for brain tumor image segmentation. *IEEE Access* (2024)

14. Wu, X., Yang, X., Li, Z., Liu, L., Xia, Y.: Multimodal brain tumor image segmentation based on densenet. *Plos one* **19**(1), e0286125 (2024)
15. Yang, W.A., Lautan, D., Weng, T.W., Lin, W.C., Kao, Y., Chen, C.C.: Global convolutional self-action module for fast brain tumor image segmentation. *IEEE Transactions on Emerging Topics in Computational Intelligence* (2024)
16. Zhu, Z., Sun, M., Qi, G., Li, Y., Gao, X., Liu, Y.: Sparse dynamic volume transunet with multi-level edge fusion for brain tumor segmentation. *Computers in Biology and Medicine* p. 108284 (2024)
17. Siyi, X., ZHANG, Y., Sixu, D. Mingwei, W., Jiangang, C., Tong, T., Qinquan, G., Chantong, L., Menghan, H., Tao, T.: Arga-unet: Advanced u-net segmentation model using residual grouped convolution and attention mechanism for brain tumor mri image segmentation. *Virtual Reality & Intelligent Hardware* **6**(3), 203-216 (2024)
18. Li, P., Li, Z., Wang, Z., Li, C., Wang, M.: mresu-net: multi-scale residual u-net-based brain tumor segmentation from multimodal mri. *Medical & biological engineering & computing* **62**(3), 641-651 (2024)
19. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16×16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020)
20. Chollet, F.: Xception: Deep learning with depthwise separable convolutions pp. 1251-1258 (2017)
21. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861* (2017)
22. Yang, L., Zhang, R.Y., Li, L., Xie, X.: Simam: A simple, parameter-free attention module for convolutional neural networks pp. 11863-11874 (2021)
23. ShaiK, N.S., Cherukuri, T.K.: Multi-level attention network: application to brain tumor classification. *Signal, Image and Video Processing* **16**(3), 817-824 (2022)
24. Ramasamy, G., Singh, T., Yuan, X.: Multi-modal semantic segmentation model using encoder based link-net architecture for brats 2020 challenge. *Procedia Computer Science* **218**, 732-740 (2023)
25. Yang, T., Lu, X., Yang, L., Yang, M., Chen, J., Zhao, H.: Application of mri image segmentation algorithm for brain tumors based on improved yolo. *Frontiers in Neuroscience* **18**, 1510175 (2025)
26. Lyu, Y., Tian, X.: Mwg-unet++: Hybrid transformer u-net model for brain tumor segmentation in mri scans. *Bioengineering* **12**(2), 140 (2025)
27. Yang, F., Wang, F., Dong, P., Wang, B.: Hca-former:hybrid convolution attention transformer for 3d medical image segmentation. *Biomedical Signal Processing and Control* **90**, 105834 (2024)
28. Jetlin, C., et al.: Pyqdcnn: Pyramid qdcnn for multi-level brain tumor classification using mri image. *Biomedical Signal Processing and Control* **100**, 107042 (2025)