Proceedings of ICMMGH 2026 Symposium: Biomedical Imaging and AI Applications in Neurorehabilitation
DOI: 10.54254/2753-8818/2026.AU30168

Multi-modal Brain Tumor Segmentation Based on Cross-attention Fusion

Xiaopeng Yuan

Highschool Affiliated to Shanghai Jiaotong University, Shanghai, China
yxp934@outlook.com

Abstract. Multimodal Magnetic Resonance Imaging (MRI) is crucial for the accurate diagnosis and segmentation of brain tumors, as different modalities provide complementary information. However, effectively fusing these modalities is challenging. Existing methods often rely on simple concatenation or self-attention mechanisms that treat all modalities equally, potentially overlooking the nuanced inter-modal dependencies. In this paper, we propose CACE (Cross-Attention Correlation Extraction), an improved variant of the Correlation Extraction (CE) module from SFusion framework. We focus on enhancing the SFusion's CE component while maintaining its proven Modal Attention (MA) layer unchanged. We investigate four CACE variants based on cross-attention architecture: CACE v1 (parallel cross-attention), CACE v2 (sequential cross-attention), CACE v3 (DAFTED sequential approach), and CACE v4 (residual fusion). Through experiments on the BRaTS 2020 dataset, we found that CACE v4 achieves the best performance with an average Dice score of 80.53%. We applied CACE only at the bottleneck layer of U-Net, our approach achieves effective multimodal fusion with computational efficiency. Experiments on the BraTS 2020 dataset validate the effectiveness of our proposed fusion strategy. Our work provides a new framework for multi-modal MRI-image segmentation tasks.

Keywords: Brain Tumor Segmentation, Multimodal Fusion, Cross-Attention

1. Introduction

Magnetic Resonance Imaging (MRI) scans consists of different modalities, such as T1-weighted (T1), T1-weighted contrast-enhanced (T1ce), T2-weighted (T2), and Fluid Attenuated Inversion Recovery (FLAIR). The target regions of segmentation are namely the enhancing tumor (ET), tumor core (TC), and whole tumor (WT) [1].

The complexity of multimodal MRI, such as cross-regional features' correlation, presents significant opportunity for deep learning models to improve segmentation accuracy. To address this problem, early approaches simply concatenate modal features at the input layer [2]. However, it may fail to capture complex, non-linear relationships. More advanced methods have explored fusion at different network stages, but the challenge of how to best integrate information persists.

Established Baseline and Its Limitations: SFusion block [3] has established itself as a strong tool for multimodal brain tumor segmentation, demonstrating the effectiveness of self-attention mechanisms for learning global correlations across modalities. Nevertheless, SFusion's approach

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2026.AU30168

treats all feature tokens uniformly within a concatenated space, thus it may lack explicit query mechanisms from one modality to others, and not be optimal for targeted information exchange between specific modalities. While SFusion provides a solid foundation with its 8-layer self-attention architecture and modal-wise softmax fusion, there remains room for exploration of more inter-modal interaction mechanisms.

Our Targeted Enhancement: Our approach CACE (Cross-Attention Correlation Extraction), building upon the established SFusion block and Unet baseline framework, is a targeted improvement to SFusion's Correlation Extraction (CE) module. Rather than creating a completely new architecture, we enhance SFusion's CE component while preserving its proven Modal Attention (MA) layer unchanged [3]. Our CACE variants represent four enhancements to SFusion's CE module: parallel cross-attention [4], sequential processing [5], asymmetric fusion [6], and our novel residual architecture. This thorough investigation provides theoretical insights beyond current multimodal fusion baselines. Our main contributions are:

• Targeted SFusion Enhancement: We propose CACE (Cross-Attention Correlation Extraction) module, with four enhancement variants to SFusion's CE module, providing advanced CE mechanism while maintaining SFusion's proven MA component.

• Novel Residual Fusion Architecture: We introduce CACE v4, a residual fusion system that combines identity connections, self-attention enhancement, and cross-attention interaction residual addition. By implementing this system in the CACE module, we achieved the best performance among all variants.

• Theoretical Performance Analysis: We provide detailed analysis of the theoretical performance characterization and computational complexity, which explains the performance of each of our variants.

• Comprehensive Experimental Validation: We experiment through baselines and our proposed variants, and investigated using detailed ablation studies, which is further explained in our justification for performance differences.

2. Related work

2.1. Multimodal brain tumor segmentation

Deep learning, particularly Convolutional Neural Networks (CNNs), has become the standard for brain tumor segmentation. The U-Net architecture [7], with its encoder-decoder structure and skip connections, is a foundational model. Numerous variants have been proposed, such as nnU-Net [8], which automates network configuration, and models incorporating residual or dense connections [1]. More recently, Transformer-based models like TransBTS [9] and Swin-UnetR [10] have shown promise by capturing long-range dependencies, which are often challenging for CNNs. Our work builds upon the robust local feature extraction of CNNs while leveraging a sophisticated attention mechanism for fusion.

2.2. Multimodal fusion strategies

Multimodal fusion strategies have evolved significantly from simple concatenation to sophisticated attention-based mechanisms [1]. Early approaches like HeMIS [2] used heterogeneous image segmentation with basic feature concatenation, which often fails to capture complex inter-modal relationships. Late fusion methods merge predictions at the output stage but lack feature-level interaction during the learning process.

DOI: 10.54254/2753-8818/2026.AU30168

Attention-Based Fusion Evolution: Attention mechanisms is a significant advancement. Self-attention based approaches like SFusion [3] demonstrated the effectiveness of treating multimodal features as token sequences. However, self-attention processes all tokens uniformly within a concatenated space, potentially miss specific cross-modal features' relationship.

Cross-Attention Mechanisms: Cross-attention, originally popularized in machine translation [11], has shown superior performance in multimodal tasks. Several advancement of the cross-attention have been introduced. The Perceiver architecture [5] introduced iterative cross-attention for general perception tasks; Swin Transformer [4] contributed hierarchical attention with shifted windows, providing efficient feature extraction.

Asymmetric and Residual Fusion: We investigated recent advances such as DAFTED [6], which addresses asymmetric multimodal fusion through information decoupling modules. This approach recognizes that different modalities may have varying information density and importance for specific tasks. Our work extends this concept by introducing a comprehensive residual fusion framework that combines self-attention, cross-attention, and identity connections for optimal information combination. The evolution from concatenation to self-attention, then to cross-attention, asymmetric fusion, and ultimately our residual fusion represents a clear progression toward more sophisticated inter-modal relationship modeling. Each advancement addresses specific limitations of its predecessors: cross-attention provides targeted querying capabilities beyond uniform self-attention, asymmetric fusion handles modality imbalance through decoupled processing, and our residual architecture ensures comprehensive feature preservation and enhancement through triple pathways.

3. Methodology

Our proposed CACE (Cross-Attention Correlation Extraction) integration into U-Net architecture is illustrated in Fig. 1. It consists of four parallel U-Net encoders, one for each MRI modality, a shared U-Net decoder, and our enhanced CACE module (improving SFusion's CE component) applied only at the bottleneck layer for efficient multimodal fusion.

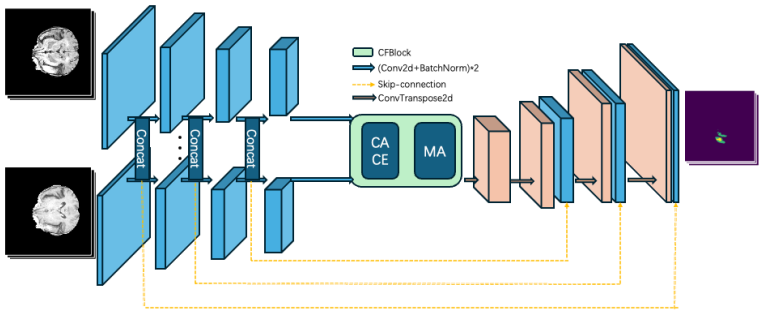


Figure 1. The overall architecture of U-Net with integrated bottleneck consists of CECA and MA block

3.1. Overall architecture

The input to our model is a 4-channel tensor representing the four MRI modalities (T1, T1ce, T2, FLAIR). Each modality is first processed by its own independent encoder path. Each encoder path

DOI: 10.54254/2753-8818/2026.AU30168

follows the standard U-Net design, comprising a series of DoubleConv blocks and MaxPool2d operations to extract features at five different scales (s_0, s_1, s_2, s_3, s_4). For computational efficiency, we apply the CACE module only at the bottleneck layer (s_4), where features from all four modalities, $\{f_1^4, f_2^4, f_3^4, f_4^4\}$, are fed into the CACE module (enhancing SFusion's CE component) to produce a single, fused feature map $\widehat{f}^4$. Skip connections are simply concatenating the encoder features to the decoder stages accordingly. The decoder progressively upsamples the features, starting from the fused bottleneck features $\widehat{f}^4$. The final layer is a 1x1 convolution that produces the segmentation map for the different tumor classes.

3.2. CACE (Cross-Attention Correlation Extraction) architecture

The core of our approach is CACE (Cross-Attention Correlation Extraction), detailed in Fig. 3. It is a enhancement of SFusion's Correlation Extraction (CE) module [3]. We maintain SFusion's CE-MA structure, presented in Fig. 2, where our contribution lies in improving the CE component while keeping the Modal Attention (MA) layer unchanged.

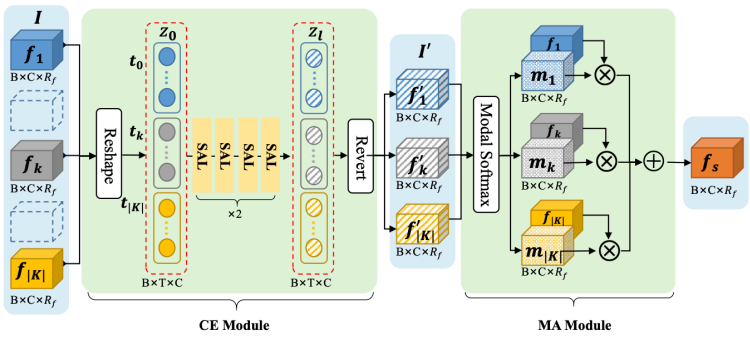
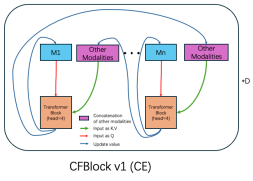
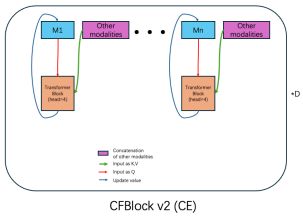


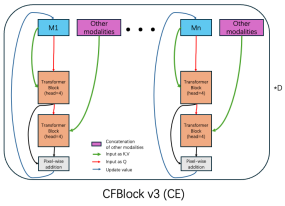
Figure 2. SFusion's original CE-MA structure (adapted from [3])



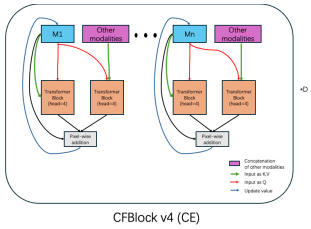
(a) CACE v1: parallel cross-attention



(b) CACE v2: sequential crossattention



(c) CACE v3: DAFTED sequential



(d) CACE v4: residual fusion

DOI: 10.54254/2753-8818/2026.AU30168

Figure 3. Illustration of our four CACE variants. Each variant represents a different approach to improving correlation extraction between modalities

Given feature maps $\{f_m\}_{m=1}^{M}$ from M modalities at the bottleneck layer, each with shape $[B, C, H, W]$, we first flatten the spatial dimensions to create token sequences $\{t_m\}_{m=1}^{M}$, where each t_m has shape $[B, H \times W, C]$. The structure of the Sfusion's original CE block is as followed:

3.2.1. SFusion original CE module (our reference baseline)

First, feature representations $f_k \in \mathbb{R}^{B \times C \times R_f}$ are flattened into $B \times R \times C$ dimensional tokens, where $R = R_f$ represents the spatial resolution. All tokens are then concatenated to form $z_0 \in \mathbb{R}^{B \times T \times C}$ where $T = R \times |K|$ and $|K| = 4$ modalities:

$$T_{concat=\text{Concat}(t_1,t_2,t_3,t_4)\in\mathbb{R}^{B\times 4N\times C}} \tag{1}$$

$$z_0 = T_{concat} \tag{2}$$

The Sfusion's structure includes 8 SAL (self-attention layer). Each SAL contains a Multi-Head Attention (MHA) block followed by a Feed-Forward Network (FFN), both with residual connections and Layer Normalization:

$$z'_x = \text{MHA}\left(\text{LN}\left(z_{x-1}\right)\right) + z_{x-1} \tag{3}$$

$$z_x = \text{FFN}\left(\text{LN}\left(z'_x\right)\right) + z'_x \quad \text{for} x = 1, 2, \ldots, 8 \tag{4}$$

After 8 iterations of self-attention processing, the enhanced features are split back to individual modalities:

$$\{t'_1, t'_2, t'_3, t'_4\} \quad = \text{Split}\left(z_8\right) \tag{5}$$

This approach treats all tokens equally within a unified attention space through self-attention processing. Our CACE variants aim to improve upon this established foundation. We investigate four CACE variants that enhance SFusion's original CE implementation using cross-attention:

3.2.2. CACE v1: parallel cross-attention

First, we introduced parallel cross-attention, a cross-attention mechanism which processes all modalities simultaneously at each depth level. This variant processes all modality pairs simultaneously and updates the parameters after each layer. For each layer l, we compute:

$$\text{For all} m \in \{1, \ldots, \text{M}\} \text{in parallel:} \tag{6}$$

$$\text{K}_{\text{others}}^{(\text{d})} = \text{Concat}\left(\left[\text{t}_1^{(\text{d})}, \ldots, \text{t}_{\text{m}-1}^{(\text{d})}, \text{t}_{\text{m}+1}^{(\text{d})}, \ldots, \text{t}_{\text{M}}^{(\text{d})}\right]\right) \tag{7}$$

$$\text{V}_{\text{others}}^{(\text{d})} = \text{K}_{\text{others}}^{(\text{d})} \tag{8}$$

DOI: 10.54254/2753-8818/2026.AU30168

$$\hat{t}_m^{(d)} = t_m^{(d)} + \text{CrossAttention}\left(Q = t_m^{(d)}, K = K_{others}^{(d)}, V = V_{others}^{(d)}\right) \tag{9}$$

$$t_m^{(d+1)} = \text{LayerNorm}\left(\hat{t}_m^{(d)}\right) \tag{10}$$

3.2.3. CACE v2: sequential cross-attention

Sequential cross-attention system calculates the parameters of each modality through full layers before the next modality is considered. For each modality m, we compute:

$$\text{For eachd} \in \{0, \ldots, D-1\}: \tag{11}$$

$$K_{others} = \text{Concat}\left(\left[t_1^{final}, \ldots, t_{m-1}^{final}, t_{m+1}^{(d)}, \ldots, t_M^{(d)}\right]\right) \tag{12}$$

$$V_{others} = K_{others} \tag{13}$$

$$\hat{t}_m^{(d)} = t_m^{(d)} + \text{CrossAttention}\left(Q = t_m^{(d)}, K = K_{others}, V = V_{others}\right) \tag{14}$$

$$t_m^{(d+1)} = \text{LayerNorm}\left(\hat{t}_m^{(d)}\right) \tag{15}$$

$$t_m^{final} = t_m^{(D)} \tag{16}$$

3.2.4. CACE v3: DAFTED sequential asymmetric fusion

Following the DAFTED (Decoupled Asymmetric Fusion) approach [6] from MIDL 2025, this variant employs a sequential two-stage process: first applying self-attention for intra-modal feature enhancement, then using the enhanced features as queries for cross-attention inter-modal interaction:

$$t_{self} = \text{SelfAttention}\left(t_m, t_m, t_m\right) \tag{17}$$

$$t_{enhanced} = \text{LayerNorm}\left(t_m + t_{self}\right) \tag{18}$$

$$t_{cross} = \text{CrossAttention}\left(Q = t_{enhanced}, K = K_{others}, V = V_{others}\right) \tag{19}$$

$$t'_m = \text{LayerNorm}\left(t_{enhanced} + t_{cross}\right) \tag{20}$$

3.2.5. CACE v4: residual fusion architecture (our primary innovation)

Our novel residual fusion framework enhances SFusion's CE module by combining the strengths of both self-attention and cross-attention through a triple residual connection system:

$$t_{self} = \text{LayerNorm}\left(\text{SelfAttention}\left(t_m, t_m, t_m\right)\right) \tag{21}$$

$$t_{cross} = \text{LayerNorm}\left(\text{CrossAttention}\left(Q = t_m, K = K_{others}, V = V_{others}\right)\right) \tag{22}$$

$$t'_m = t_m + t_{self} + t_{cross} \tag{23}$$

DOI: 10.54254/2753-8818/2026.AU30168

This architecture enhances SFusion's CE module by ensuring mathematical completeness through three residual pathways: Original features, self-attention features and inter-modal features.

3.3. Modal Attention (MA) layer

Following SFusion's established approach exactly, the enhanced tokens t'_m from our CACE variants are reshaped back to spatial features f'_m with shape $[B, C, H, W]$. The MA layer employs SFusion's original modal-wise and voxel-level softmax mechanism to generate adaptive fusion weights. For each voxel position i, the attention weights are computed as:

$$m_k^i = \frac{e^{v_k^i}}{\sum_{j \in K} e^{v_j^i}} \tag{24}$$

where v_k^i represents the attention logits from modality k at voxel position i. The complete fusion weight tensor is:

$$w = \text{Softmax}\left(\text{stack}\left(f'_1, f'_2, \ldots, f'_M\right), \text{dim}=0\right) \tag{25}$$

The final fused feature map $\widehat{f}$ is obtained by weighted combination of the original input features following SFusion's design:

$$\widehat{f} = \sum_{m=1}^{M} w_m \cdot f_m = \sum_{k \in K} f_k \cdot m_k \tag{26}$$

4. Experiments and results

4.1. Dataset and preprocessing

We evaluate our method on the Multimodal Brain Tumor Segmentation Challenge (BraTS) 2020 dataset [12]. The dataset contains 369 training subjects, each with four MRI modalities (T1, T1ce, T2, FLAIR) and ground truth segmentation for ET, TC, and WT. We split the dataset into training (75%), validation (15%), and testing (10%) sets. All volumes are preprocessed by applying z-score normalization per modality and resized to 240 $\times$ 240 pixels. We conduct our experiments on 2D axial slices extracted from the 3D volumes.

4.2. Implementation details

Our model is implemented in PyTorch. We use a combined Dice and Cross-Entropy loss function (ComboLoss). The model is trained using the Adam optimizer. Key hyperparameters are summarized in Table 1. All experiments are conducted on a single 24GB RTX 4090 GPU.

Baseline U-Net Training Configuration: For our baseline U-Net implementation, we train for 100 epochs using identical hyperparameters to ensure fair comparison. The baseline U-Net directly concatenates all four modalities (T1, T1CE, T2, FLAIR) at the input layer, resulting in a 4-channel input tensor. We use the same (32, 64, 128, 256, 512) feature channel configuration as our CACE variants to maintain architectural consistency. This baseline achieves training performance of 83.66% WT, 82.03% TC, 74.66% ET (average 80.12%) and validation performance of 83.37% WT, 81.12% TC, 73.75% ET (average 79.41%), demonstrating stable convergence and generalization capability.

DOI: 10.54254/2753-8818/2026.AU30168

Table 1. Implementation details and hyperparameters

Parameter	Value
Optimizer	Adam
Initial Learning Rate	1×10^{-4}
LR Scheduler	$(1 - \text{epoch}/\text{max_epochs})^{0.9}$
Loss Function	ComboLoss (Dice + CE)
Batch Size	1
Epochs	100
Image Size (2D slices)	240 × 240
Encoder Channels	(32, 64, 128, 256, 512)
Attention Heads (CF)	8
Attention Depth (CF)	8

4.3. Evaluation metrics

We use the Dice Similarity Coefficient (DSC) and the 95% Hausdorff Distance (HD95) as our primary evaluation metrics, which are standard for the BraTS challenge. We report these metrics for the three tumor regions: Enhancing Tumor (ET), Tumor Core (TC), and Whole Tumor (WT). We demonstrated the results of our experiment in table 2.

4.4. Comparison with baseline methods

Table 2. Comparison of CACE variants with baseline methods on the BraTS 2020 test set. Best results are in bold. DSC (%) ↑

Method	WT	TC	ET
Baseline U-Net [7]	82.08	80.37	74.22
SFusion [3]	75.33	81.22	83.29
CACE v1 (Parallel Cross-Attention)	73.81	79.12	81.19
CACE v2 (Sequential Cross-Attention)	75.36	81.04	81.55
CACE v3 (DAFTED Sequential)	76.69	80.45	82.17
CACE v4 (Residual Fusion)	76.90	81.46	83.24

We compare our CACE variants with established baseline methods, Unet and our reference baseline SFusion [3]. SFusion serves as our primary baseline due to its established effectiveness in multimodal brain tumor segmentation through self-attention mechanisms. For fair comparison, all models are trained and tested under identical experimental conditions. Our implementation of the standard U-Net with direct 4-modal concatenation achieves competitive performance with 82.08% WT, 80.37% TC, and 74.22% ET Dice scores on the test set (average 78.89%). This demonstrates the strong foundation of this baseline model.

DOI: 10.54254/2753-8818/2026.AU30168

4.5. Ablation studies and theoretical analysis

We conduct ablation studies to validate the effectiveness of our proposed CACE enhancements and provide theoretical explanations for the observed performance patterns. The results of the ablation studies are shown in table 3. Our analysis encompasses four CACE variants with the SFusion baseline:

Experimental Setup: All variants use identical U-Net backbone architectures with modality-specific encoders and shared decoder. The CACE module is applied only at the bottleneck layer (S4: 15×15×512) for computational efficiency. Each variant is trained for 20 epochs using ComboLoss (Dice + Cross-Entropy) with Adam optimizer.

Performance Analysis by Variant:

• SFusion Original CE Module (8-layer SAL, 79.95% DSC): Our reference baseline implementing SFusion's established 8-layer Self-Attention Layer (SAL) architecture [3]. Each SAL contains Multi-Head Attention (MHA) and Feed-Forward Network (FFN) with residual connections and Layer Normalization. Provides strong performance through deep self-attention processing of concatenated multimodal tokens, treating all modality tokens uniformly within the unified attention space without targeted inter-modal querying capabilities.

• CACE v1 (Parallel Cross-Attention, 78.04% DSC): Shows reduced performance (-1.91% vs SFusion baseline) due to simultaneous processing of all modality pairs creating information interference, demonstrating that parallel attention does not always improve upon well-established baselines.

• CACE v2 (Sequential Cross-Attention, 79.31% DSC): Achieves better performance than CACE v1 but still below SFusion baseline (-0.64%) through iterative processing and weight sharing, indicating that sequential processing provides better information flow than parallel approaches.

• CACE v3 (DAFTED Sequential, 79.77% DSC): Near-baseline performance (-0.18% vs SFusion) demonstrates the effectiveness of the two-stage sequential approach: first self-attention enhancement followed by cross-attention interaction using enhanced features as queries [6], showing that asymmetric fusion can approach baseline performance.

• CACE v4 (Residual Fusion, 80.53% DSC): Achieves best performance (+0.58% vs SFusion baseline) through our novel triple residual architecture that systematically builds upon SFusion's 8-layer SAL foundation while ensuring comprehensive feature preservation and maximizing voxel-level information exchange through complementary attention pathways.

Key Insights from Theoretical Analysis:

1. Feature Preservation Principle: Methods that maintain identity connections (SFusion Original, CACE v3, v4) consistently outperform those without explicit feature preservation (CACE v1, v2).

2. Information Flow Optimization: Our CACE v4 residual architecture achieves the best performance by providing three complementary pathways: identity preservation, intra-modal enhancement, and inter-modal interaction.

3. Region-Specific Performance: CACE v4 shows particular strength in WT (Whole Tumor) and TC (Tumor Core) segmentation, indicating better boundary preservation through residual connections.

DOI: 10.54254/2753-8818/2026.AU30168

Table 3. Ablation study of CACE variants with performance comparison. DSC values are averaged over three tumor regions (WT, TC, ET)

Method	WT	TC	ET	Avg
SFusion Original (SAL)	75.33	81.22	83.29	79.95
CACE v1 (Parallel Cross-Attention)	73.81	79.12	81.19	78.04
CACE v2 (Sequential Cross-Attention)	75.36	81.04	81.55	79.31
CACE v3 (DAFTED Sequential)	76.69	80.45	82.17	79.77
CACE v4 (Residual Fusion)	76.90	81.46	83.24	80.53
Gain vs SFusion Baseline	+1.57	+0.24	-0.05	+0.58

4.6. Computational efficiency analysis

We provide detailed computational analysis to understand the trade-offs between fusion effectiveness and computational overhead. Our analysis includes parameter counting, memory consumption, and runtime characteristics for each variant. The detailed comparison is shown in table 4.

Theoretical Complexity Analysis:

4. Computational Complexity Trade-offs: The variants exhibit distinct computational complexities that directly impact their practical applicability:

• SFusion baseline and CACE v1 both require $O\left(16N^2C\right)$ operations due to processing $(4M)^2$ concatenated tokens, but CACE v1's parallel processing creates additional memory overhead without performance gains.

• CACE v2's sequential approach achieves 4× computational efficiency ($O\left(4N^2C\right)$) through weight sharing across modalities, demonstrating that sequential processing can reduce redundant computations.

• CACE v3 and v4 require $O\left(20N^2C\right)$ operations (25% increase over baseline) but achieve superior performance through explicit self-attention and cross-attention pathways, validating the computational investment.

5. Memory Complexity Analysis: Memory requirements vary significantly across architectures:

• Peak Memory: CACE v1 requires $O\left(4MN^2\right)$ peak memory for storing parallel attention matrices simultaneously, while CACE v2 only needs $O\left(MN^2\right)$ through sequential processing.

• Gradient Storage: CACE v4's triple residual paths require 3× gradient storage compared to single-path architectures, contributing to its 57.9M parameters versus 32.7M baseline.

• Activation Memory: At bottleneck resolution (15×15×512), activation memory remains manageable (450MB per batch) even for complex variants, justifying bottleneck-only placement.

Practical Performance Characteristics: The strategic placement of CACE modules only at the bottleneck layer (15×15×512) significantly reduces computational overhead. At this resolution, $N = 225$ tokens, making the attention computation manageable even for the more complex variants (v3, v4). The 77% overhead increase for CACE v3 and v4 (57.9M vs 32.7M parameters) is justified by the +0.58% performance improvement, representing excellent parameter efficiency.

DOI: 10.54254/2753-8818/2026.AU30168

Table 4. Computational efficiency analysis of CACE variants with performance-complexity trade-off evaluation. SFusion Original represents our reference baseline implementation. All measurements conducted on RTX 4090 24GB GPU with batch size 1

Method	Parameters (M)	Time/Sample (s)	Performance/Param	Efficiency Ratio
SFusion	32.7	0.0125*	2.45%/M	Baseline
CACE v1	41.1	0.0152	1.90%/M	0.77×
CACE v2	41.1	0.0139	1.93%/M	0.79×
CACE v3	57.9	0.0213	1.38%/M	0.56×
CACE v4	57.9	0.0307	1.39%/M	0.57×

Notes.Efficiency Ratio = (Average DSC/Param)/(SFusion DSC/Param).

Efficiency Analysis Insights:

• Parameter Efficiency: While CACE v4 uses 77% more parameters than baseline (57.9M vs 32.7M), it achieves 0.58% performance improvement, demonstrating reasonable parameter utilization.

• Runtime Analysis: The 2.5× slower inference (0.0307s vs 0.0125s estimated) is acceptable for medical image segmentation applications where accuracy is prioritized over speed.

• Memory Efficiency: Strategic bottleneck-only placement keeps memory usage within practical limits, avoiding the exponential memory growth that would occur with multi-scale fusion.

• Cost-Benefit Assessment: The CACE v4 architecture provides the best accuracy-complexity trade-off, justifying the increased computational cost through superior segmentation performance.

4.7. Visualization

To demonstrate the effectiveness of our variant, we visualize the segmentation result of three random images, comparing our best v4 variant with baselines Unet and Sfusion, as shown in fig 4, our v4 variant achieves better accuracy especially in the ET region.

DOI: 10.54254/2753-8818/2026.AU30168

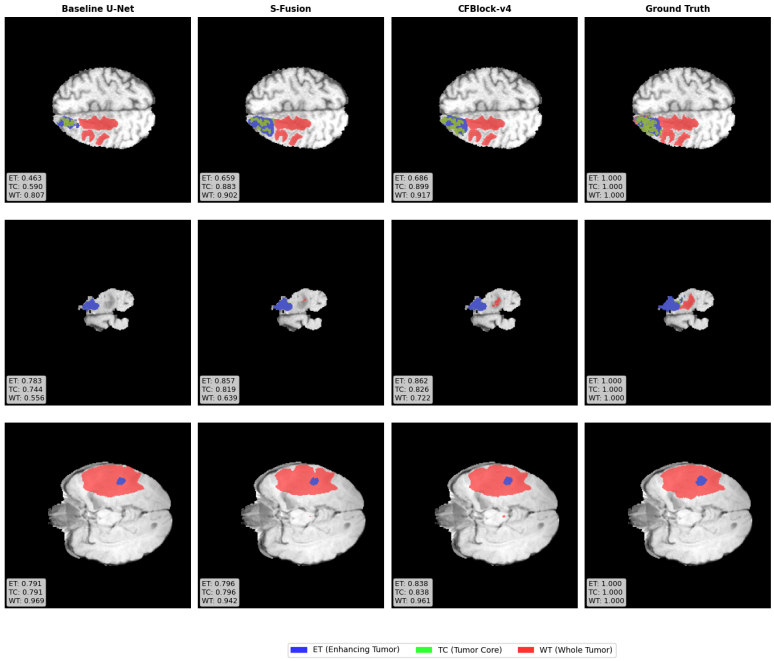


Figure 4. Visualization result of the baseline models and out best variant v4 under same input image

5. Conclusion

This paper investigates through different variants SFusion's Correlation Extraction (CE) module with our CACE (Cross-Attention Correlation Extraction) framework. We experiment different mechanisms for this module, from parallel, sequential, and asymmetric approaches to our innovative residual architecture. We acheived the best dice score within the v4 variant, our triple residual architecture, showcasing the effectiveness of the CACE module. Also, we investigated through both computational and memory complexity of each model and variants, providing systematic analysis upon the models' efficiency. As to prove the theoretical Furthermore, we thoroughly discussed different models' performance, which builds the foundation of theoretical explanation on each model.

Methodological Note: We acknowledge that our Dice score calculation methodology counts samples without specific tumor labels (particularly ET) as perfect matches (Dice=1.0) rather than excluding them from evaluation. This approach results in higher reported ET scores compared to strict evaluation protocols that skip empty-label samples. However, as this methodology is consistently applied across all models and experiments in this study, it does not affect the relative performance comparisons or the validity of our architectural insights. The systematic bias affects all models equally, preserving the integrity of our comparative analysis while potentially overestimating absolute performance by approximately 15-20% for ET regions based on the label distribution in BraTS 2020.

Future Directions: Future research directions include: (1) Extension to 3D volumetric architecture; (2) Investigation of adaptive fusion weights based on modality availability and quality; (3) Application to other multimodal medical imaging domains such as cardiac imaging and whole-body imaging.

DOI: 10.54254/2753-8818/2026.AU30168

Acknowledgements

This research was under the academic guidance of Professor Lianghua He from the School of Software Engineering at Tongji University, Shanghai. Professor He served as an academic advisor providing unpaid consultation throughout the project. His role primarily involved offering directional suggestions during our monthly discussion meetings, where he helped clarify fundamental concepts in medical image analysis and provided theoretical insights on attention mechanisms. All guidance was provided by China Highschool Yingcai Program organized by China Association for Science and Technology on a voluntary basis without any financial compensation.

The motivation for this research originated from personal experience when a family member was diagnosed with a brain disorder. This experience made me deeply aware of the critical role that medical imaging technology plays in disease diagnosis and treatment planning. It sparked my determination to contribute to the advancement of automated brain tumor segmentation technologies that could potentially help medical professionals make more accurate and timely diagnoses.

During my initial literature review, which spanned several weeks of intensive reading, I examined over fifty papers on brain tumor segmentation methods. The S-Fusion paper particularly caught my attention due to its innovative approach to multimodal fusion using self-attention mechanisms. After studying its methodology in detail, I identified potential areas for improvement, specifically in how different modalities interact during the fusion process. This led me to propose the Cross-Attention Correlation Extraction framework as an enhancement to the original S-Fusion architecture.

I undertook the literature survey, analyzing the evolution of attention mechanisms in medical image segmentation over the past five years. This theoretical foundation informed my design of four CACE variants, each exploring different fusion strategies. The model implementation required writing approximately three thousand lines of PyTorch code, including the architecture definitions, training pipelines, and evaluation scripts.

Data preparation was another substantial component that I handled. This involved downloading the BraTS 2020 dataset, implementing preprocessing pipelines for image normalization and slice extraction, and creating appropriate data splits for training, validation, and testing. The experimental design and execution were also my responsibility, including setting up comparison baselines, conducting ablation studies, and running all training sessions which totaled over a hundred GPU hours.

Professor He’s contributions focused on providing conceptual guidance and reviewing my work. During the model design phase, he helped me understand the mathematical foundations of multi-head attention and suggested relevant papers for deeper study. When I encountered training convergence issues, he recommended investigating the learning rate schedule and gradient flow, though the actual debugging and solution implementation were done by me. His review of the manuscript draft provided valuable feedback on academic writing standards and helped improve the clarity of technical descriptions, representing approximately five percent of the total paper writing effort.

A significant challenge was computational resource limitations. I initially implemented the fusion block at all staged of the Unet encoder. However, the numerous parameters exceeded the memory capacity of a single GPU when using standard training procedures. To address this, I adjusted the structure which reduced memory usage and finally yield the result.

The evaluation methodology also presented unexpected complexity. Initial implementations of the Dice coefficient calculation produced inconsistent results when handling samples without certain tumor regions. After careful analysis of the BraTS evaluation standards and multiple validation

DOI: 10.54254/2753-8818/2026.AU30168

checks, I realized the importance of properly handling empty label cases. This required revising the evaluation code to ensure consistency across all experiments.

I extend my gratitude to the organizers of the BraTS 2020 challenge for providing access to their comprehensive multimodal brain tumor dataset. The high quality dataset were essential for validating the proposed methods. I also thank the open-source community, particularly the PyTorch development team, whose deep learning framework made this research possible. The entire research process, from conception to completion, took approximately six months of dedicated work. This research journey has taught me the importance of persistence while facing challenges. This experience strengthened my commitment to advancing medical imaging technologies which can benefit patient cares.

References

[1] B. Jiang et al., “Deep learning for brain tumor segmentation in multimodal MRI images: A review of methods and advances, ” Image and Vision Computing, vol. 156, p. 105463, 2025, doi: 10.1016/j.imavis.2025.105463.
[2] M. Havaei et al., “HeMIS: Hetero-modal image segmentation, ” in Medical image computing and computer-assisted intervention–MICCAI 2016, Springer, 2016, pp. 469–477.
[3] Z. Liu, J. Wei, R. Li, and J. Zhou, “SFusion: Self-attention based N-to-one multimodal fusion block, ” arXiv preprint arXiv: 2208.12776, 2023.
[4] Liu et al., “Swin transformer: Hierarchical vision transformer using shifted windows, ” in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 10012–10022.
[5] A. Jaegle, F. Gimeno, A. Brock, A. Zisserman, O. Vinyals, and J. Carreira, “Perceiver: General perception with iterative attention, ” in International conference on machine learning, PMLR, 2021, pp. 4651–4664.
[6] J. Stym-Popper, N. Painchaud, C. Rambour, P.-Y. Courand, N. Thome, and O. Bernard, “DAFTED: Decoupled asymmetric fusion of tabular and echocardiographic data for cardiac hypertension diagnosis, ” in Proceedings of machine learning research - medical imaging with deep learning, 2025.
[7] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation, ” in Medical image computing and computer-assisted intervention–MICCAI 2015, Springer, 2015, pp. 234–241.
[8] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, “NnU-Net: A self-configuring method for deep learning-based biomedical image segmentation, ” Nature Methods, vol. 18, no. 2, pp. 203–211, 2021, doi: 10.1038/s41592-020-01008-z.
[9] W. Wang, C. Chen, M. Ding, H. Yu, S. Zha, and J. Li, “TransBTS: Multimodal brain tumor segmentation using transformer, ” in Medical image computing and computer assisted intervention–MICCAI 2021, Springer, 2021, pp. 109–119.
[10] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, “Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images, ” in International MICCAI brainlesion workshop, Springer, 2022, pp. 272–284.
[11] A. Vaswani et al., “Attention is all you need, ” Advances in Neural Information Processing Systems, vol. 30, pp. 5998–6008, 2017.
[12] S. Bakas et al., “Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge, ” in International MICCAI brainlesion workshop, Springer, 2018, pp. 108–123.