TransSea: Hybrid CNN–Transformer With Semantic Awareness for 3-D Brain Tumor Segmentation

Yu Liu, *Member, IEEE*, Yize Ma, Zhiqin Zhu, Juan Cheng, *Member, IEEE*, and Xun Chen, *Senior Member, IEEE*

***Abstract*— Accurate segmentation of brain tumors in multimodal magnetic resonance imaging (MRI) plays a crucial role in clinical quantitative assessments, diagnostic processes, and the planning of therapeutic strategies. Both convolutional neural networks (CNNs) with strong local information extraction capacities and Transformers with excellent global representation capacities have achieved remarkable performance in medical image segmentation. However, considering the inherent semantic disparities between local and global features, effectively combining convolutions and Transformers presents a significant challenge in medical image segmentation. To address this issue, through integrating the merits of these two paradigms in a well-designed encoder–decoder architecture, we propose a hybrid CNN–Transformer network with semantic awareness, named TransSea, for an accurate 3-D brain tumor segmentation task. Our network incorporates a semantic mutual attention (SMA) module at the encoding stage, seamlessly integrating global and local features. Furthermore, our design includes a multiscale semantic guidance (SG) module that introduces semantic priors in the encoder through semantic supervision, enabling focused segmentation in relevant areas. In the decoding process, a semantic integration (SI) module is presented to further integrate various feature mappings from the encoder and semantic priors, thereby enhancing the propagation of semantic information and achieving semantically aware querying. Extensive experiments on two brain tumor datasets, BraTS2020 and BraTS2021, demonstrate that our model significantly outperforms existing state-of-the-art methods. The source code of the proposed method will be made available at https://github.com/yuliu316316.**

***Index Terms*— Brain tumor segmentation, convolutional neural networks (CNNs), multimodal magnetic resonance imaging (MRI), semantic guidance (SG), Transformer.**

I. Introduction

BRAIN tumors are aberrantly growing cell clusters within the brain that pose substantial threats to patients' lives and well-being. Accurate and reproducible quantification, alongside morphological analysis of these tumors, is crucial for diagnosis, treatment planning, and outcome assessment [1]. Brain tumor segmentation aims to accurately identify distinct tumor regions in medical images [2]. Among various imaging techniques, magnetic resonance imaging (MRI) stands as the primary diagnostic method for brain tumors. It aids in spatial mapping and analyzing disease progression. This is achieved through the integration of diverse 3-D MRI modalities, such as FLAIR, T1-weighted, contrast-enhanced T1-weighted, and T2-weighted sequences [3]. Specifically, brain tumor segmentation involves identifying key components: enhancing tumor (ET), necrosis and nonenhancing tumor (NCR/NET), and peritumoral edema/infiltration (ED). Fig. 1 demonstrates the application of various 3-D MRI modalities in brain tumor segmentation and the corresponding 3-D segmentation visualizations. Notably, the ground truth (GT) segmentation label, approved by expert neuroradiologists, is shown in Fig. 1(e), in which green, yellow, and red indicate ED, ET, and NCR/NET regions, respectively [4]. The differences in cross-modal features are evident across Fig. 1(b)–(e).

With the advent of deep learning, numerous convolution-based methods have been developed for segmenting specific targets in medical images. U-shaped encoder–decoder architectures, such as U-Net [5] and fully convolutional networks (FCNs) [6], are dominating this field along with their variants, such as U-Net++ [7], Attention U-Net [8], 3-D U-Net [9], and V-Net [10]. These models have achieved significant success in brain tumor segmentation, demonstrating the potent capability of convolutional neural networks (CNNs) in learning semantic information. Nevertheless, the limited receptive field of CNNs restricts their effectiveness in capturing extensive spatial dependencies, which is crucial for image segmentation. To overcome this limitation, Transformer-based models [11], [12], known for their self-attention mechanisms and global contextual information capturing ability, have been integrated into brain tumor segmentation research [13]. However, the fixed input size of standard Vision Transformers (ViTs) [12] entails high computational costs for pixel-level dense predictions in semantic segmentation. The Swin Transformer [14] addresses this with its hierarchical structure that yields CNN-like feature maps, offering substantial benefits in semantic segmentation tasks. Nevertheless, the Swin Transformer still faces challenges, such as low local inductive bias, requiring extensive training data for effective visual representation [15].

Manuscript received 29 January 2024; revised 10 May 2024; accepted 19 May 2024. Date of publication 12 June 2024; date of current version 24 June 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62176081, Grant U23A20294, and Grant 62171176; and in part by the Science and Technology Innovation Key Research and Development Program of Chongqing under Grant CSTB2022TIAD-KPX0039. The Associate Editor coordinating the review process was Dr. Mohammad Forouzanfar. *(Corresponding author: Zhiqin Zhu.)*

Yu Liu, Yize Ma, and Juan Cheng are with the Department of Biomedical Engineering and Anhui Province Key Laboratory of Measuring Theory and Precision Instrument, Hefei University of Technology, Hefei 230009, China (e-mail: yuliu@hfut.edu.cn; mayize@mail.hfut.edu.cn; chengjuan@hfut.edu.cn).

Zhiqin Zhu is with the College of Automation, Chongqing University of Posts and Telecommunications, Chongqing 400065, China (e-mail: zhuzq@cqupt.edu.cn).

Xun Chen is with the Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei 230027, China (e-mail: xunchen@ustc.edu.cn).

Digital Object Identifier 10.1109/TIM.2024.3413130

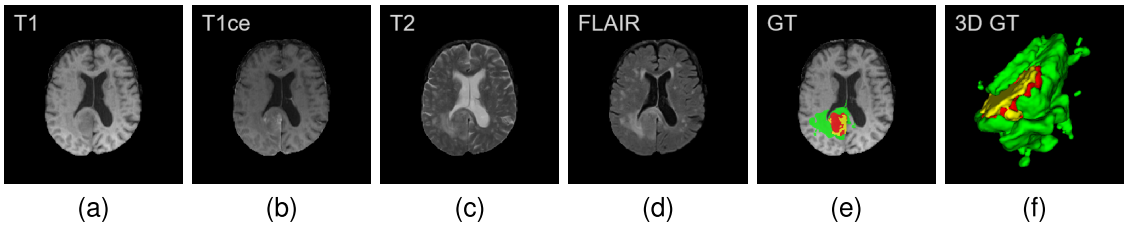

Fig. 1. Examples of multimodal MRI for brain tumor segmentation and 3-D visualization of GT. (a) T1 modality. (b) T1ce modality. (c) T2 modality. (d) FLAIR modality. (e) GT segmentation label provided by domain experts (the green, yellow, and red represent edema (ED), ET, and NCR/NET, respectively). (f) GT segmentation 3-D visualization.

In medical image segmentation, simultaneously learning global semantic information and detailed local features is essential. Inspired by CNNs, which have powerful local information extraction capabilities, and Transformers with excellent global and long-distance representation abilities, some concurrent studies, such as CvT [16], PVT [17], TransUNet [18], and Segtran [19], have attempted to combine convolution with Transformers in different ways, aiming to integrate the strengths of both CNNs and Transformer paradigms. For CvT and PVT, the linear projection of self-attention in Transformers is replaced with convolution. TransUNet and Segtran connect Transformer modules behind multiple stacked convolutional layers. However, due to the semantic ambiguity between the local features of convolution and the global features of Transformers, the simple combination of existing models [18], [19] fails to integrate the advantages of both effectively. In addition, many existing methods [17], [20], [21] mainly focus on extracting and fusing deep semantic features. They often neglect the semantic context available during the image encoding phase and the diverse semantic information in different regions, both crucial for accurate segmentation. While some methods [22], [23] introduce semantic mapping as a prior and integrate it into feature representation to enhance segmentation performance, they still fall short in fully exploiting the available semantic knowledge from segmentation networks. Such methods exhibit a deficiency in the thorough processing of semantic information across various regions, thus limiting potential improvements in segmentation performance [24], [25]. Finally, due to the quadratic computational complexity of feature numbers, Transformers are computationally intensive, making the complexity of combining Transformers and convolution overwhelming. Therefore, exploring the incorporation of semantic knowledge and query-aware guidance in designing brain tumor segmentation models using convolution and Transformers remains an area with significant uncharted potential.

To address the above issues, this article introduces TransSea, a 3-D brain tumor segmentation network based on an encoder–decoder architecture. This network is designed to more effectively utilize multimodal information and semantic context. It employs CNNs for meticulous local feature extraction and Transformers for capturing comprehensive contextual data, effectively bridging the semantic gap between local and global understanding in medical imaging. Moreover, it incorporates query-based semantic priors, seamlessly integrating them with the decoding features to achieve a heightened level of semantic perception. This novel architecture not only addresses the typical challenges found in brain tumor segmentation but also introduces a methodical integration of semantic awareness at every stage of the segmentation process, setting it apart from existing methodologies. Overall, the main contributions of this article include the following.

1) We propose TransSea, a novel semantic-aware hybrid model, designed for the automatic segmentation of brain tumors in multimodal MRI images. TransSea effectively captures both local details and global semantic information, showcasing its semantic sensitivity. This method brings the following three technical contributions.
2) We present a semantic mutual attention (SMA) module into the encoder, combining the Swin Transformer's capabilities with the advantages of depth-spatial separable convolution (DSConv). In particular, it embeds semantic awareness into the encoder via a semantically guided attention mechanism, allowing for more effective semantic representation.
3) We present a semantic guidance (SG) module, aimed at providing early semantic supervision and generating query values for guiding feature fusion during the decoding stage. This module strategically enhances the encoder's focus on regions crucial for segmentation, leading to improved feature transmission across various scales.
4) We present a semantic integration (SI) module for multilevel and cascaded integration. Leveraging semantically aware attention, this module dynamically integrates query-based, semantically guided features with those of the decoder, thereby effectively utilizing feature-level semantic information.

The remainder of this article is organized as follows. Section II discusses the related work. In Section III, the proposed method is depicted in detail. Experimental results and discussion are given in Section IV. Finally, Section V concludes this whole article.

II. Related Work

A. Brain Tumor Segmentation Methods

Accurate brain tumor segmentation typically precedes critical measurements, such as tumor volume and response assessment in neuro-oncology (RANO) metrics, such as the maximum bidimensional diameters' product. With more accurate segmentation results, the corresponding measurements

will be more accurate accordingly, which is of great significance to the relevant response assessment tasks [26], [27], [28].

Early medical image segmentation systems were primarily based on traditional image segmentation algorithms [29], involving methods based on edge detection, threshold-based methods, and region-based methods. However, medical images, especially MRI images, often exhibit characteristics such as low contrast, complex textures, and areas with blurred boundaries, which limit the effectiveness and applicability of these types of image segmentation algorithms.

In recent years, CNNs [30], [31] have become indispensable in supervised learning across various computer vision tasks, particularly in the field of brain tumor segmentation. Medical image segmentation networks based on the U-Net [5] architecture have achieved remarkable success. Variants of U-Net, such as UNet++ [32] and Res-UNet [33], have significantly enhanced performance. Metlek and Çetıner [34] introduced a convolution-based hybrid model that applies convolutions specifically to regions of interest identified across different modalities. Montaha et al. [35] utilized a 2-D U-Net architecture automated for 3-D MRI scans, highlighting the model's effectiveness across various MRI sequences. However, these methods, based on 2-D slice images, fail to fully utilize the interslice information. To address this issue, Çiçek et al. [9] introduced the 3-D U-Net, the first 3-D segmentation network for medical imaging. Similar lightweight 3-D CNN-based models include DMFNet proposed by Wang et al. [36], which achieved high segmentation performance on the BraTS 2018 dataset. Liu et al. [26], [37] demonstrated the effectiveness of multimodality MRI image fusion in enhancing segmentation accuracy. Wang et al.'s CPNet [36] efficiently aggregated spatial information through fully separable convolutions (separated in both spatial and depth dimensions) for contextual inference. Further advancing CNN-based approaches, Aboussaleh et al. [38] introduced an efficient U-Net architecture incorporating three different encoders. By fusing feature maps from each encoder and integrating them with an attention mechanism, this advanced approach effectively segments various types of brain tumors. However, influenced by the locality of the convolution receptive field, these methods cannot capture global or long-range contextual interaction information and spatial dependency information, which is crucial for medical image segmentation.

Transformer (ViT) [12] demonstrated the applicability and effectiveness of Transformer technology in computer vision downstream tasks such as classification, detection, and segmentation. Subsequently, many Transformer-based segmentation methods have achieved state-of-the-art performance in medical image segmentation. However, the original multihead self-attention (MHSA) mechanism of ViT encounters high computational costs when processing high-resolution inputs. To mitigate this, local self-attention Transformer methods, such as the Swin Transformer [14] and CSwin Transformer [39], have been proposed and proven effective. Nevertheless, in contrast to CNNs, Transformers inherently lack sensitivity to local structures, particularly when segmenting images directly into patch-based tokens. This intrinsic limitation highlights a fundamental divergence in their approach to image processing.

B. CNN–Transformer Hybrid Networks

Inspired by the formidable capability of CNNs in extracting local information and the advantage of Transformers in global and long-distance representations, a trend has emerged in the field of visual Transformers: the integration of convolutional operations with self-attention mechanisms, harnessing the strengths of both. This hybrid type of visual Transformer, the CNN–Transformer architecture, effectively combines the best of both paradigms and excels in visual tasks. For instance, Chen et al.'s TransUNet [18] integrated CNNs and Transformers in its architecture, with CNNs handling feature extraction and Transformation, while Transformers managed global context encoding. However, this direct approach of segmenting images into Transformer tokens overlooked the local structure of 3-D volumetric data. To address this problem, Wang et al. [13] proposed TransBTS, the first model to apply Transformers to 3-D CNNs for 3-D MRI brain tumor segmentation. Other models, such as Swin UNETR [40], have integrated CNNs and Transformers in their building blocks [41], [42], [43], demonstrating the remarkable effectiveness of this hybrid model across various visual tasks. However, existing methods based on a sequential connection structure of encoders cannot efficiently extract and fuse global–local features.

To obtain more accurate segmentation results, fully utilizing semantic context plays a crucial role in brain tumor segmentation. Zhang et al. [44] proposed a context encoding module for capturing global semantic context. Recently, Jin et al. [45], [46] have focused on capturing and integrating semantic-level context information as well as image-level context information, realized in specially designed decoders. Moreover, Feng et al. [47] and Liang et al. [48] integrate advanced features such as deep supervision and attention mechanisms. Notably, these works capture semantic context based on features extracted after the encoding stage, rather than the encoder's ability to capture semantic features, leading to a loss of semantic information during the encoding phase. Jain et al. [49] introduced a framework to capture semantic information in the encoding stage. However, the above methods have not fully utilized the extracted semantic information.

III. Proposed Method

A. Overview

As depicted in Fig. 2, the proposed TransSea network architecture, based on an encoder–decoder framework, comprises primarily an SMA module, an SG module, and an SI module. These modules are designed to tackle not only the traditional challenges of semantic segmentation but also to seamlessly integrate local and global contextual information in a manner unprecedented by previous models. The encoder is constructed with 3-D DSConv, Swin Transformer, and a comprehensive semantic attention (CSA) mechanism. Its primary function is to extract and fuse global–local features while fully leveraging deep semantic features and downsampling input images from

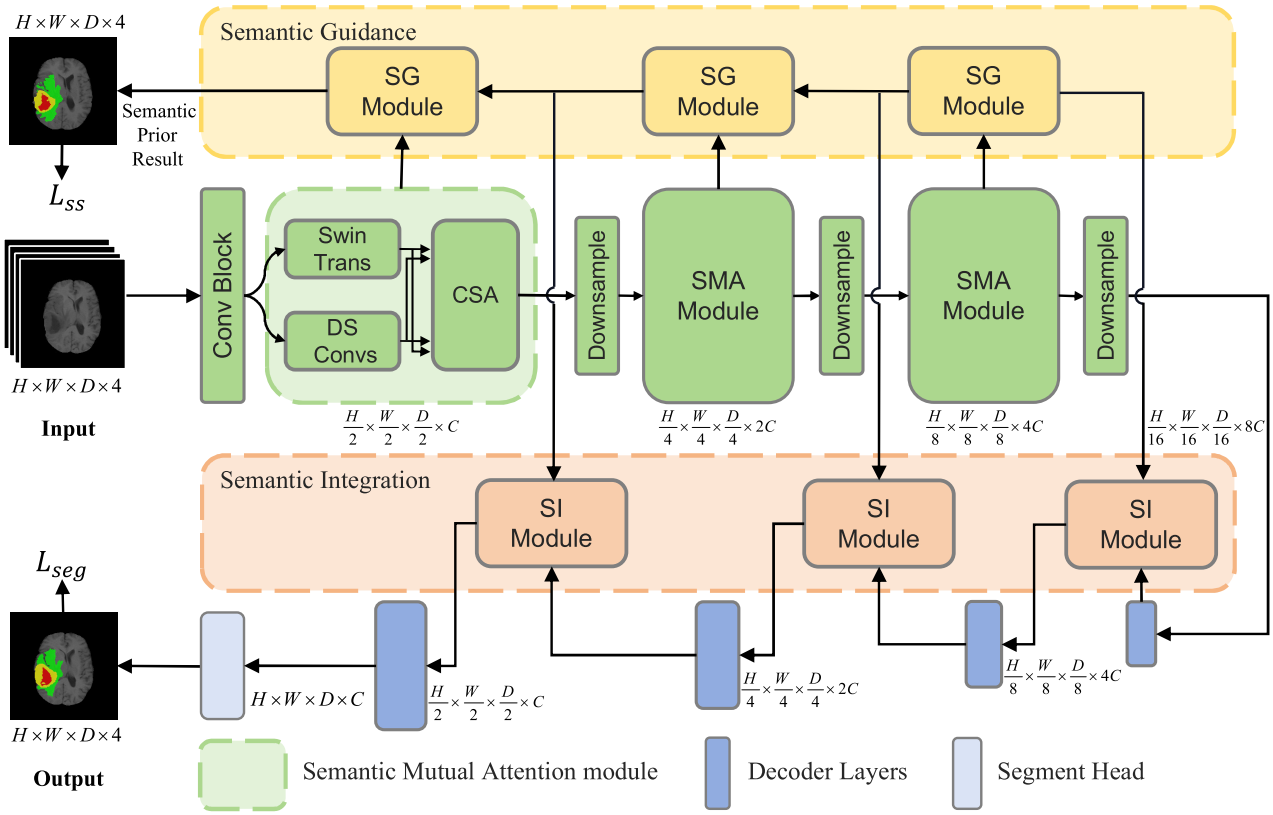

Fig. 2. Our TransSea network is architecturally grounded in an encoder–decoder framework, comprising three main modules: the SMA module, the SG module, and the SI module. The SMA module is constructed with 3-D DSConv, Swin Transformer, and a CSA mechanism, aimed at extracting and integrating global–local features from source images of different modalities. Subsequently, the SG module is crucial for early semantic supervisory learning, facilitating the generation of semantic prior results. Finally, the SI module adeptly integrates multiscale SG features with decoding features.

multimodal MRI scans, including FLAIR, T1, T1ce, and T2 modalities. In addition, we introduce an auxiliary semantic learning branch incorporating a 3-D SG module. The semantic priors from all stages are aggregated through the SG module to efficiently capture tumor-related semantic information during the encoder's training phase. Concurrently, following the general architecture of U-Net-like networks [5], [50], we employ the SI module to connect upstream SG features to the fused attention decoding module via skip connections, building a multiscale and richer integration of semantic features. In contrast to the TransSea encoder, the decoder comprises upsampling blocks and a segmentation head, primarily responsible for mapping the low-resolution feature maps, aggregated from the encoder and SI module, to pixel-level predictions. This novel architecture not only addresses the typical challenges found in brain tumor segmentation but also introduces a methodical integration of semantic awareness at every stage of the segmentation process, setting it apart from existing methodologies. Table I systematically summarizes the output sizes of the main layers in our architecture, detailing the dimensions of all variables throughout each processing stage.

B. SMA Module

The intricate interplay between local features and global representations, a subject of extensive study in the long history of visual descriptors, is crucial. To fully leverage both local and global features, we design a parallel module, named SMA module, as depicted in Fig. 3. This module simultaneously extracts local features through convolution blocks and global features through Transformer blocks, aiming to enrich the semantic information of the extracted features. Recognizing the complementary nature of these two styles of features, we introduce the CSA mechanism. Inspired by the success of cross-attention [51], [52], [53], this mechanism effectively fuses local and global features with significant semantic differences. Following this, the output features from this module are downsampled and input into the decoder, where they are fused with the output features from the SG module to achieve improved segmentation performance.

1) Transformer Branch: To reduce the computational complexity of the model when extracting global features, we employ the 3-D Swin Transformer [54] in the architecture, as illustrated in Fig. 3(a), instead of the ViT. Compared to the standard Transformer, the Swin Transformer introduces a hierarchical design that partitions the input image into nonoverlapping image blocks and leverages window-based multihead self-attention (W-MSA) to reduce computational load. Due to the sliding-window segmentation operation performed by W-MSA, the cropped blocks do not overlap, and there is a lack of effective information interaction between the windows. To further enhance the model's performance, it incorporates a unique shifted W-MSA (SW-MSA) in the

TABLE I
OUTPUT SIZES OF THE MAIN LAYERS IN OUR PROPOSED ARCHITECTURE, WHICH INCLUDES THE DIMENSIONS OF VARIABLES USED AT EACH STAGE OF PROCESSING

Layer	Used Variables	Output Size
Input	-	$(H \times W \times D \times 4)$
Conv	$\boldsymbol{Z}$	$(\frac{H}{2} \times \frac{W}{2} \times \frac{D}{2} \times C)$
SMA[1]+DS[2]	$\boldsymbol{Z}_{hybrid}$	$[(\frac{H}{4} \times \frac{W}{4} \times \frac{D}{4} \times 2C), (\frac{H}{8} \times \frac{W}{8} \times \frac{D}{8} \times 4C), (\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16} \times 8C)]$
SG[3]	$\boldsymbol{y}$	$[(\frac{H}{8} \times \frac{W}{8} \times \frac{D}{8} \times 4C), (\frac{H}{4} \times \frac{W}{4} \times \frac{D}{4} \times 2C), (\frac{H}{2} \times \frac{W}{2} \times \frac{D}{2} \times C)]$
SI[4]	$\boldsymbol{X}^i_{\text{out}}$	$[(\frac{H}{8} \times \frac{W}{8} \times \frac{D}{8} \times 4C), (\frac{H}{4} \times \frac{W}{4} \times \frac{D}{4} \times 2C), (\frac{H}{8} \times \frac{W}{8} \times \frac{D}{8} \times 4C), (\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16} \times 8C)]$
Decoder	-	$[(\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16} \times 8C), (\frac{H}{8} \times \frac{W}{8} \times \frac{D}{8} \times 4C), (\frac{H}{4} \times \frac{W}{4} \times \frac{D}{4} \times 2C), (\frac{H}{2} \times \frac{W}{2} \times \frac{D}{2} \times C)]$
Output	-	$(H \times W \times D \times 4)$

[1]Semantic Mutual Attention Module, [2]Downsampling, [3]Semantic Guidance Module, [4]Semantic Integration Module.

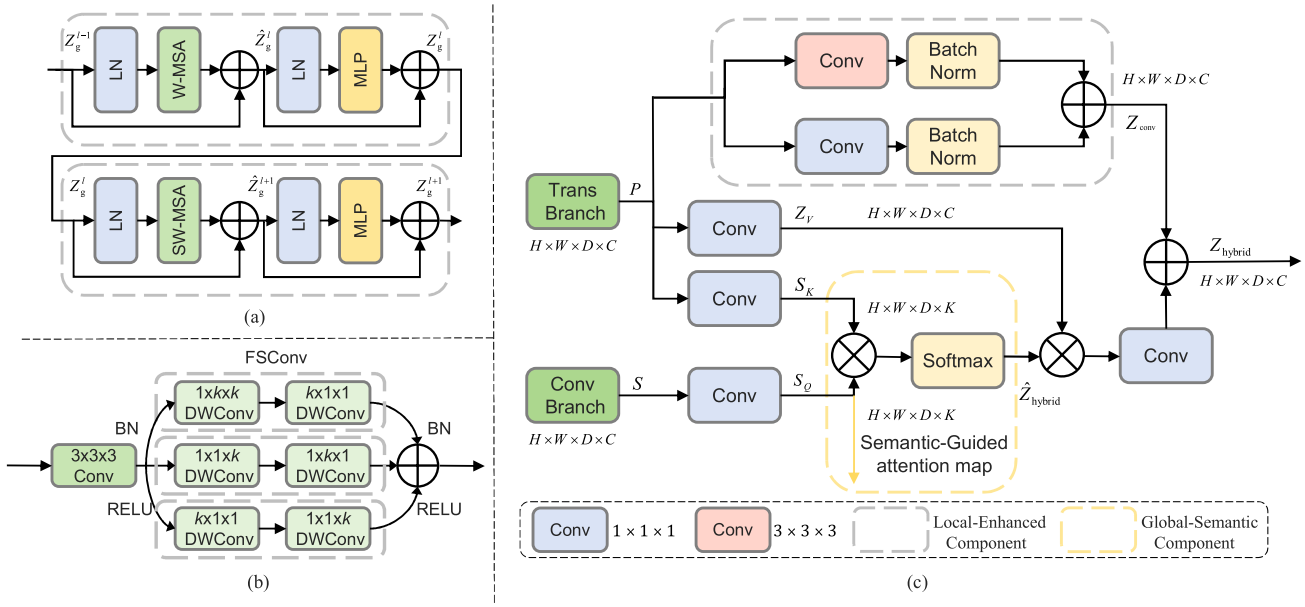

Fig. 3. Workflow of the proposed SMA module. (a) Successive Swin Transformer layers. (b) DSConv. (c) CSA.

next consecutive layers, enhancing the modeling of long-range dependencies and, consequently, improving the ability to capture contextual information. By using the 3-D W-MSA and 3-D SW-MSA methods, consecutive 3-D Swin Transformer layers can be represented as follows:

$$\hat{\boldsymbol{Z}}^l_{\text{glo}} = \text{3DW} - \text{MSA}\left(\text{LN}\left(\boldsymbol{Z}^{l-1}_{\text{glo}}\right)\right) + \boldsymbol{Z}^{l-1}_{\text{glo}} \tag{1}$$

$$\boldsymbol{Z}^l_{\text{glo}} = \text{MLP}\left(\text{LN}\left(\hat{\boldsymbol{Z}}^l_{\text{glo}}\right)\right) + \hat{\boldsymbol{Z}}^l_{\text{glo}} \tag{2}$$

$$\hat{\boldsymbol{Z}}^{l+1}_{\text{glo}} = \text{3DSW} - \text{MSA}\left(\text{LN}\left(\boldsymbol{Z}^l_{\text{glo}}\right)\right) + \boldsymbol{Z}^l_{\text{glo}} \tag{3}$$

$$\boldsymbol{Z}^{l+1}_{\text{glo}} = \text{MLP}\left(\text{LN}\left(\hat{\boldsymbol{Z}}^{l+1}_{\text{g0}}\right)\right) + \hat{\boldsymbol{Z}}^{l+1}_{\text{glo}} \tag{4}$$

where $\hat{\boldsymbol{Z}}^l_{\text{glo}}$ and $\boldsymbol{Z}^l_{\text{glo}}$ represent the output global features of the 3-D (S)W-MSA module and MLP module in the lth 3-D Swin Transformer block, respectively. MLP and LN denote the multilayer perceptron module and layer normalization, respectively. According to [14], we employ 3-D cyclic shifting for efficient batch computation of shifted windowing. The W-MSA module and the SW-MSA module primarily consist of self-attention mechanisms and trainable relative position encoding. The overall computation is defined as follows:

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{d_k}} + \boldsymbol{B}\right)\boldsymbol{V} \tag{5}$$

where $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ represent the query, key, and value matrices, respectively; d_k is the $\boldsymbol{Q}/\boldsymbol{K}$ dimension; and $\boldsymbol{B}$ represents the deviation in relative position information. Through this batch computation approach, 3-D SW-MSA maintains efficient computational complexity consistent with 3-D W-MSA.

2) CNN Branch: Similar to [55], [56], and [57], we have devised efficient DSConv, decomposing standard convolutions into three asymmetric spatial convolutions. For a $k \times k \times k$ convolution, we employ a $1 \times k \times k$ convolution followed by a $k \times 1 \times 1$ convolution, referred to as spatial separable convolution. This technique reduces computational complexity while preserving the same receptive field size as standard convolutions. Moreover, each spatial separable convolution utilizes depthwise convolution [58], [59] with dimensions,

further enhancing computational efficiency. Fig. 3(b) provides an overview of the complete DSConvs structure, where each branch effectively captures various aspects of volumetric spatial information. These branches combine depth-wise semantic information with spatial features, resulting in a holistic understanding of the input volume. In tasks such as brain tumor segmentation, different tissue types exhibit diverse semantic information in 3-D space. This design allows the network to process features at varying depths or spatial levels of input data, facilitating a more profound comprehension of semantic context in images and the extraction of local features from different regions.

3) Comprehensive Semantic Attention: Considering the semantic differences between global and local features, how to utilize the complementarity of these two styles of features and fuse them is a significant challenge. To address this issue, we propose the CSA mechanism, as illustrated in Fig. 3(c), comprising primary input and secondary input, which interactively couple local features with global representations in a continuous manner. With input features $\boldsymbol{Z} \in R^{H\times W\times D\times C}$, the CSA can be expressed as

$$\boldsymbol{S}_Q = f_1(\boldsymbol{S}), \quad \boldsymbol{S}_K = f_2(\boldsymbol{P}), \ \boldsymbol{Z}_v = f_3(\boldsymbol{P}) \tag{6}$$

$$\boldsymbol{Z}_{\text{conv}} = \text{BN}(f_4(\boldsymbol{P}) + f_5(\boldsymbol{P})) \tag{7}$$

$$\boldsymbol{W}_{\text{SegAttn}} = \text{Softmax}\left(\frac{\boldsymbol{S}_Q \boldsymbol{S}_K{}^{\top}}{\sqrt{d}}\right) \tag{8}$$

$$\hat{\boldsymbol{Z}}_{\text{hybrid}} = \boldsymbol{W}_{\text{SegAttn}} \boldsymbol{Z}_v \tag{9}$$

$$\boldsymbol{Z}_{\text{hybrid}} = f_6\left(\hat{\boldsymbol{Z}}_{\text{hybrid}}\right) + \boldsymbol{Z}_{\text{conv}} \tag{10}$$

where $\boldsymbol{P}$ and $\boldsymbol{S}$ are the primary and secondary input features, respectively; $f_1(\cdot)$, $f_2(\cdot)$, $f_3(\cdot)$, $f_4(\cdot)$, $f_5(\cdot)$, and $f_6(\cdot)$ are linear projections with 3D convolution; d is the number of channels in $\boldsymbol{S}/\boldsymbol{P}$; $\boldsymbol{S}_Q \in \mathbb{R}^{(H\times W\times D)\times K}$, $\boldsymbol{S}_K \in \mathbb{R}^{(H\times W\times D)\times K}$, and $\boldsymbol{Z}_v \in \mathbb{R}^{(H\times W\times D)\times C}$ represent semantic query (SQ), semantic key (SK), and feature value (FV), respectively, with K being equal to the number of segmentation classes; $\boldsymbol{W}_{\text{SegAttn}}$ and $\boldsymbol{Z}_{\text{hybrid}}$ represent the semantic attention matrix and the final output, respectively. Thus, the attention matrix $\boldsymbol{W}_{\text{SegAttn}}$ reflects the importance of primary features to secondary features, subsequently obtaining the hybrid features $\hat{\boldsymbol{Z}}_{\text{hybrid}}$ through formula (9). For efficient backpropagation during training, we finally obtain the output $\boldsymbol{Z}_{\text{hybrid}}$ by summing $\boldsymbol{Z}_{\text{conv}}$ and $\hat{\boldsymbol{Z}}_{\text{hybrid}}$.

In CSA, the global semantic component is responsible for capturing semantic context within our encoder. It updates features from the input through segmentation scores, offering guidance for effective supervision in semantic modeling and providing semantic-guided attention maps. It splits features P and S from the previous module into three projections: SQ, SK, and FV, enriched with local contextual information. The attention matrix is obtained by multiplying query and key in self-attention [11], providing early semantic supervision learning for the query to guide the activation of features of interest and mask irrelevant ones. The local-enhanced component employs two parallel different convolution blocks, followed by batch normalization operations, to enhance the extraction of local context. Global and local features are sequentially input into the attention as primary and secondary branches, respectively, and then concatenated along the channel dimension.

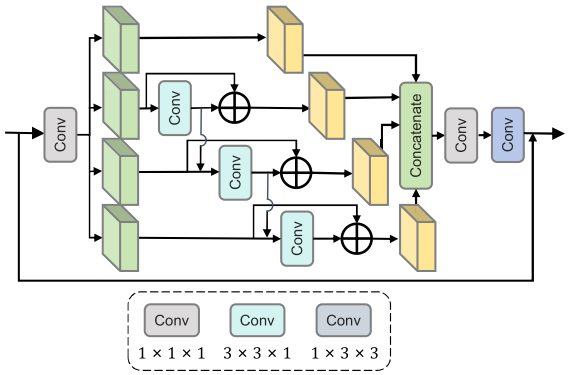

Fig. 4. Architecture of the SG module. This diagram illustrates the SG module, designed to provide early semantic supervision and generate query values that guide feature fusion during the decoding process. It consists of a series of 3-D convolutions that strategically process features from different stages of the encoder, enhancing focus on regions crucial for accurate segmentation.

C. SG Module

Recently, Jain et al. [49] demonstrated that introducing semantic priors into established hierarchical encoders can enhance the performance of image semantic segmentation. 3DPSWIN [48] has proven that semantic supervision assists models in effectively masking irrelevant noise areas while updating weights. In segmentation tasks, it is crucial for the model to focus more on the relevance between the pixels to be segmented and other pixels. In the encoder, the global semantic component of the SMA module is used to separate semantic information from features, generating semantic priors, which are then used to update features based on these semantic prior maps. This provides early semantic supervision learning for query to guide the activation of interesting features and mask irrelevant queries. Therefore, as shown in Fig. 4, we utilize the SG module to pool features and semantic prior maps from different stages of the encoder.

Unlike some concurrent works [60], [61] that enhance multiscale capabilities by utilizing features of different resolutions, our proposed method's multiscale refers to multiple available receptive fields on a finer granularity level. It also uses skip connections to gather information from a broader range of pixels, thereby avoiding the grid effect. The SG module consists of a series of 3-D convolutions, skip connections, and summation operations that pool features and semantic-guided attention maps from different stages, making it an efficient and intuitive segmentation decoder for our purposes. The SG module can be represented by the following formula:

$$\boldsymbol{x}_1^{\text{in}}, \boldsymbol{x}_2^{\text{in}}, \boldsymbol{x}_3^{\text{in}}, \boldsymbol{x}_4^{\text{in}} = S(\boldsymbol{x}) \tag{11}$$

$$\boldsymbol{x}_r^{\text{out}} = \begin{cases} \boldsymbol{x}_r^{\text{in}}, & r = 1 \\ \text{Conv}\left(\boldsymbol{x}_r^{\text{in}}\right) + \boldsymbol{x}_r^{\text{in}}, & r = 2 \\ \text{Conv}\left(\boldsymbol{x}_r^{\text{in}} + \boldsymbol{x}_{r-1}^{\text{in}}\right) + \boldsymbol{x}_r^{\text{in}}, & r = 3, 4 \end{cases} \tag{12}$$

$$\boldsymbol{y} = A\left(S\left(C\left(\boldsymbol{x}_1^{\text{out}}, \boldsymbol{x}_2^{\text{out}}, \boldsymbol{x}_3^{\text{out}}, \boldsymbol{x}_4^{\text{out}}\right)\right)\right) + \boldsymbol{x} \tag{13}$$

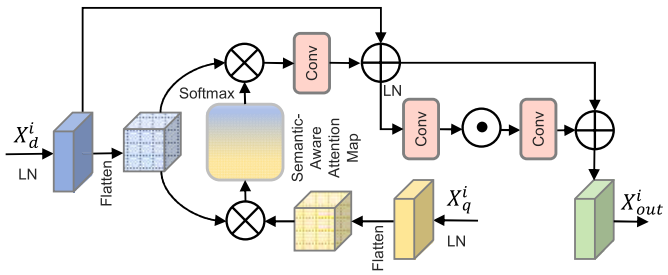

Fig. 5. Architecture of the SI module. This figure shows the SI module, which performs multilevel and cascaded integration of semantic features. It utilizes a series of operations, including softmax normalization, cross-modal similarity calculations, and semantic-aware attention mappings to dynamically integrate query-based, semantically guided features from the SG module with those from the decoder.

where $\boldsymbol{y}$ represents the output feature map of the SG module and $\boldsymbol{x}$ represents the input feature map of the SG module. The input $\boldsymbol{x}$ is uniformly divided into four groups $\boldsymbol{x}_1^{\text{in}}$, $\boldsymbol{x}_2^{\text{in}}$, $\boldsymbol{x}_3^{\text{in}}$, and $\boldsymbol{x}_4^{\text{in}}$ by channels: $\boldsymbol{x}_1^{\text{out}}$, $\boldsymbol{x}_2^{\text{out}}$, $\boldsymbol{x}_3^{\text{out}}$, and $\boldsymbol{x}_4^{\text{out}}$ represent the output of each group after feature extraction. G represents a $1 \times 3 \times 3$ convolution, H represents a $1 \times 1 \times 1$ convolution, C represents the concatenation operation, and *Conv* represents a $3 \times 3 \times 1$ convolution.

Specifically, in the process of processing the input $\boldsymbol{x}$ through convolution, we first evenly divide the channels into four groups. The first group, $\boldsymbol{x}_1$, employs a standard convolutional kernel to extract image features, resulting in $\boldsymbol{x}_1^{\text{out}}$. The remaining $s-1$ subgroups use convolution to extract layer features relevant to brain tumors, with each layer utilizing skip connections to obtain $\boldsymbol{x}_2^{\text{out}}$, $\boldsymbol{x}_3^{\text{out}}$, and $\boldsymbol{x}_4^{\text{out}}$. Subsequently, feature maps from all groups are concatenated and sent through convolution to exchange information between channels. Finally, convolution extracts spatial feature information from the image from another perspective, yielding the output $\boldsymbol{y}$. In the SG module, the splits are processed in a multiscale manner, aiding in the extraction of both global information and local information. To better integrate information from different scales, we connect all the splits and pass them through convolution. The split and connection strategy enables convolution to process features more efficiently. To reduce the number of parameters, we omit the first split's convolution, which can also be seen as a form of feature reuse. Consequently, the output of the SG module can capture feature information from different scales of receptive fields.

D. SI Module

In our pursuit of enhancing image feature mapping through the incorporation of semantic priors, we must take into account the semantic disparities between semantic-guided features and decoder features. To address this issue, we introduce the SI module, as depicted in Fig. 5. This module serves as a bridge between the semantic-guided branch and the decoder, as illustrated in Fig. 2. Its primary function is to facilitate the creation of multiscale and cascaded semantic feature integration. Considering the potential representations of the decoder, denoted as $\boldsymbol{X}_d^i$, and the output features of the SG module, denoted as $\boldsymbol{X}_q^i$ (where i represents different layers, specifically $i = 0, 1, 2$), we combine these features to construct multiscale semantic priors. This allows us to fully leverage semantic information across various resolution levels.

In traditional self-attention mechanisms [11], [12], the time and memory complexity of key–query dot-product interactions increases quadratically with the spatial resolution of the input. To mitigate this issue, we draw inspiration from the Restormer strategy [62] and employ a low-computation-cost cross-channel cross-covariance calculation. The employed method effectively generates an attention map that implicitly encodes global context.

Specifically, the SI module processes image features through cross-modal similarity calculations and produces a semantic-aware attention map. First, we aggregate per-pixel cross-channel context using convolutional layers. Subsequently, we employ deep convolutional encoding to capture spatial context between channels, with an emphasis on local context. This transformation projects $\boldsymbol{X}_d^i$ into key and value representations, which contain rich local context information. Likewise, $\boldsymbol{X}_q^i$ from the SG module is transformed into projections of the same dimension, which serve as queries for attending to semantic-aware features. By reshaping the projections of queries and keys and performing dot-product interactions, we generate a semantic-aware map that represents the relationship between $\boldsymbol{X}_d^i$ and $\boldsymbol{X}_q^i$. To further enhance these features, we apply a feedforward network (FN) to each pixel location, ensuring that the entire module fully leverages contextual information to enrich the features. Finally, this module outputs its result as input to the subsequent enhanced network decoder. As a result, our SI module, which includes the semantic-aware attention map, can be detailed as follows:

$$T^i = \text{Softmax}\left(U_k V_k\left(\boldsymbol{X}_d^i\right) \times U_q V_q\left(\boldsymbol{X}_q^i\right)/\sqrt{C}\right) \tag{14}$$

$$X_{\text{out}}^i = \text{FN}\left(U_v V_v\left(\boldsymbol{X}_d^i\right) \times T^i + \boldsymbol{X}_d^i\right) \tag{15}$$

where $U(\cdot)$ represents a $1 \times 1 \times 1$ point convolution, $V(\cdot)$ is a $3 \times 3 \times 3$ depth convolution, C denotes the number of feature channels, $T^i \in \mathbb{R}^{C \times C}$ represents the semantic-aware attention map, and $\boldsymbol{X}_{\text{out}}^i$ is the final improved feature mapping of the ith SI module.

E. Loss Function

During model training, we calculate the total loss by summing the semantic supervision loss $\mathcal{L}_{\text{ss}}$ and segmentation loss $\mathcal{L}_{\text{seg}}$. $\mathcal{L}_{\text{seg}}$ determined based on the primary prediction from the decoder of TransSea. Following [12], [31], and [70], we utilize a cross-entropy loss and Dice loss as the segmentation loss. For the semantic supervision loss, $\mathcal{L}_{\text{ss}}$ is computed on the semantic prior prediction of the SG branch, as shown in Fig. 2. Unlike $\mathcal{L}_{\text{seg}}$, $\mathcal{L}_{\text{ss}}$ comprises solely cross-entropy loss, which hastens model convergence. Consequently, we can formulate the proposed loss function as follows:

$$\mathcal{L}_{\text{seg}} = 1 - \frac{2}{J}\sum_{j=1}^{J}\frac{\sum_{i=1}^{I} G_{i,j} Y_{i,j}}{\sum_{i=1}^{I} G_{i,j}^2 + \sum_{i=1}^{I} Y_{i,j}^2} - \frac{1}{I}\sum_{i=1}^{I}\sum_{j=1}^{J} G_{i,j} \log Y_{i,j} \tag{16}$$

Algorithm 1 Training Procedure of TransSea

Input: Training data $D = \{X^k, G^k\}_k^M$. Initializing the hyper-parameters.
Start training:
for numbers of training epochs **do**
  for numbers of iteration times **do**
    Step1, TransSea forward based on the input patch. Calculate the predicted results and semantic-prior prediction: $\{Y^k, S^k\}$
    Step2, Calculate the cost function from the numbers of batch size (m):
    $L(Y, S, G)) = \frac{1}{m}\sum_{k=1}^{m}(\mathcal{L}_{seg}(Y^k, G^k) + \lambda_s \mathcal{L}_{ss}(S^k, G^k))$
    Step3, Update the parameter θ_d of SGD optimization: $\theta_d = \theta_d - \eta \cdot \nabla_{\theta_d} L(\bar{B})$
  end for
end for
Output: Trained model parameters θ_d

$$\mathcal{L}_{\text{ss}} = -\frac{1}{I}\sum_{i=1}^{I}\sum_{j=1}^{J} G_{i,j} \log S_{i,j} \tag{17}$$

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{seg}} + \lambda_s \mathcal{L}_{\text{ss}} \tag{18}$$

where I denotes the number of voxels; J denotes the number of categories; $Y_{i,j}$ and $G_{i,j}$, respectively, denote the one-hot encoded prediction and GT of category j at voxel i; $S_{i,j}$ denotes the probability of semantic prior prediction for category j at voxel i; and λ_s is the weighting factor used to balance the semantic supervision loss and segmentation loss. The training procedure of our proposed TransSea network is summarized in Algorithm 1.

IV. Experiments

A. Experimental Settings

1) Datasets: Our research utilizes the BraTS2020 and BraTS2021 datasets from the Brain Tumor Segmentation (BraTS) Challenge [76] for training and testing our proposed model. The BraTS Challenge, spearheaded by the International Society for Medical Image Computing and Computer-Assisted Intervention (MICCAI), benchmarks the forefront algorithms in brain tumor segmentation. Comprising multiparametric MRI scans from diverse clinical institutions, the BraTS datasets are rigorously vetted by expert neuroradiologists and include precise pixel-level segmentation labels. It includes a training set, a validation set, and a test set. Since only the segmentation labels (i.e., GT labels) of the training set are publicly available, we further divided the training dataset into training, validation, and testing subsets in an 8:1:1 ratio for our experiments, which is consistent with many previous studies in this field [38], [77], [78], [79]. Details about the datasets used are presented in Table II. The BraTS2020 and BraTS2021 datasets encompass 369 and 1251 meticulously annotated samples, respectively. Each includes quadruple modality MRI scans (Flair, T1, T1ce, and T2), meticulously labeled by experts. Labels span four categories: background, NCR/NET, ED, and ET, with evaluations focusing on three tumor regions: whole tumor (WT), tumor core (TC), and ET. This research utilized two prevalent metrics for assessing performance in medical image segmentation: the Dice coefficient and the 95% Hausdorff distance (HD95).

TABLE II
Description of the Used Datasets, Including Their Sources, Types, Formats, Sizes, Sequences, Tasks, Timepoints, and the Number of Images

Datasets	BraTS2021	BraTS2020
Datasets Source	Kaggle	
Image Type	3D Brain MRI	
Image Format	NIfTI	
Image Size	240 × 240 × 155	
Sequences	FLAIR, T1, T1ce, T2, Seg	
Tasks	Segmentation	
Timepoint	Pre-operative	
Total Number of Images	1251	369
Total Number of Training Images	1000	295
Total Number of Validation Images	125	37
Total Number of Testing Images	126	37

TABLE III
Detailed Hyperparameter Setup for the Proposed Model

Hyperparameters	Value
No. of epochs	300
Optimizer	SGD
Momentum	0.9
Maximum Learning Rate	0.004 (initial)
Minimum Learning Rate	0.002 (after decay)
Learning Rate Decay	Cosine Annealing
Warm-up epochs	10
Batch size	2
λ_s	0.5
Activation function	ReLU
Loss function	Cross-Entropy + Dice

2) Data Preparation and Augmentation: In this study, extensive preprocessing is applied to the multimodal MRI data to optimize network training and inference. Initially, MRI images from four modalities (FLAIR, T1, T1ce, and T2) are merged into a unified four-channel 3-D voxel dataset. Each channel undergoes standard z-score normalization, harmonizing image contrast discrepancies across modalities. A specialized mask is employed to identify and normalize nonbackground voxels, while background voxels (zero values in all modalities) are preserved unaltered. In our data augmentation strategy, we reduce the dimensions of MRI images from the original size of 240 × 240 × 155 to a cropped size of 160 × 160 × 128 by focusing on nonzero voxels. We also incorporate rotations at fixed intervals of 90°, 180°, and 270°, which maintain the original quality of the MRI scans by avoiding resampling. To better simulate the inherent noise found in MRI data and to bolster the robustness of the model, Gaussian noise is introduced into the dataset.

TABLE IV
OBJECTIVE EVALUATION RESULTS OF DIFFERENT BRAIN TUMOR SEGMENTATION METHODS ON THE BRATS2020 BENCHMARK

Method	Year	Dice (%) ↑				Hausdorff 95% (mm) ↓			
		Avg.	ET	TC	WT	Avg.	ET	TC	WT
3D U-Net [9]	2016	78.65	72.86	77.94	85.15	20.978	31.690	18.740	12.498
Att-Unet [63]	2018	80.42	74.18	80.72	86.38	17.395	25.357	15.284	11.544
U-Net++ [32]	2020	85.06	79.83	85.57	89.77	5.370	4.328	5.483	6.299
TransBTS [13]	2021	83.08	78.08	81.02	90.14	12.883	19.697	12.788	6.164
Swin-BTS [64]	2022	82.24	77.36	80.30	89.06	17.060	26.840	15.780	8.560
VT-UNet [65]	2022	83.46	77.64	82.48	90.28	11.558	18.837	6.348	9.489
NestedFormer [66]	2022	86.16	80.05	86.40	92.03	5.051	5.269	5.316	4.567
ACMINet [67]	2023	85.48	81.13	84.70	90.61	10.193	17.500	8.630	4.450
ADHDC-Net [68]	2023	83.77	78.01	83.31	89.99	14.803	29.340	9.820	5.250
3DUV-NetR+ [69]	2024	85.48	**81.70**	82.80	91.95	4.900	**3.800**	6.000	4.900
DAUnet [47]	2024	83.80	78.60	83.00	89.80	14.267	27.600	9.800	5.400
Proposed	2024	**86.32**	79.71	**86.89**	**92.36**	**4.892**	5.896	**4.789**	**3.992**

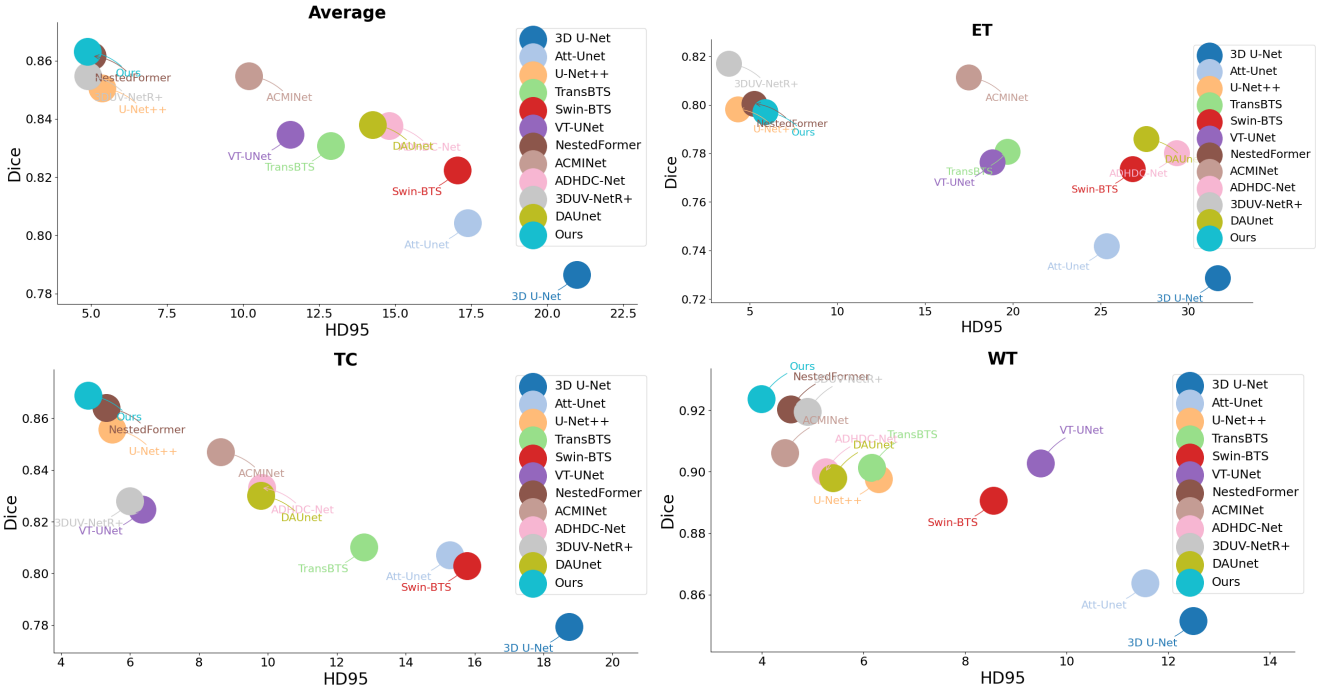

Fig. 6. Scatter plots with regard to the performance of different brain tumor segmentation methods on the BraTS2020 benchmark.

3) Implementation Details: The model was developed within the PyTorch framework, with training conducted on an NVIDIA RTX A6000 GPU. A stochastic gradient descent (SGD) optimizer was utilized, starting with a learning rate of 0.004, adjusted via a cosine annealing scheduler, with a ten-iteration warm-up, and tapering from a base rate of 0.004 to 0.002. The batch size was established at 2, with a total of 300 iterations. The hyperparameter values used in our model are presented in Table III.

B. Comparison With State-of-the-Art Methods

To validate the effectiveness of TransSea, we conduct comparisons with several state-of-the-art segmentation methods, rigorously tested on the BraTS2020 and BraTS2021 benchmark datasets. These methods encompass 3-D CNN-based networks [9], [32], [63], [67], [70], [71], Transformer-based networks [65], [66], and hybrid networks combining Transformer and CNN architectures [13], [40], [48], [64], [73], [74]. Evaluation results for different methods on the BraTS2020 and BraTS2021 benchmarks are presented in Tables IV and V, respectively. The best-performing values are indicated in bold. To offer a more intuitive comparison, Figs. 6 and 7 illustrate scatter plots regarding the performance of different models. The horizontal axis represents the HD95 value, while the vertical axis indicates the Dice score. Therefore, points closer to the top-left corner signify better performance.

TABLE V
OBJECTIVE EVALUATION RESULTS OF DIFFERENT BRAIN TUMOR SEGMENTATION METHODS ON THE BRATS2021 BENCHMARK

Method	Year	Dice (%) ↑				Hausdorff 95% (mm) ↓			
		Avg.	ET	TC	WT	Avg.	ET	TC	WT
3D U-Net [9]	2016	80.13	76.20	76.17	88.02	19.003	25.481	21.565	9.965
Att-Unet [63]	2018	83.64	79.60	91.59	89.74	14.045	19.365	14.684	8.085
DMFNet [71]	2019	84.91	80.66	83.35	90.74	15.927	27.949	13.427	6.405
TransBTS [13]	2021	85.03	81.17	83.49	90.45	11.951	18.942	10.141	6.772
VT-UNet [65]	2022	88.07	85.59	87.41	91.20	7.520	6.230	6.590	10.030
Swin UNETR [72]	2022	88.97	85.80	88.50	92.60	5.206	6.016	3.770	5.831
SegTransVAE [73]	2022	89.53	85.48	**92.60**	90.52	4.100	**2.890**	5.840	**3.570**
3D PSwinBTS [48]	2022	87.32	82.62	86.72	92.64	10.784	17.531	11.084	3.738
CKD-TransBTS [74]	2023	88.39	84.76	88.07	92.33	3.927	3.160	4.390	4.230
DAUnet [47]	2024	83.97	77.60	84.40	89.90	9.233	14.400	6.600	6.700
MAT [75]	2024	90.06	85.05	91.91	93.21	4.767	3.610	3.560	7.130
Proposed	2024	**90.84**	**86.49**	92.33	**93.70**	**3.207**	2.964	**3.024**	3.632

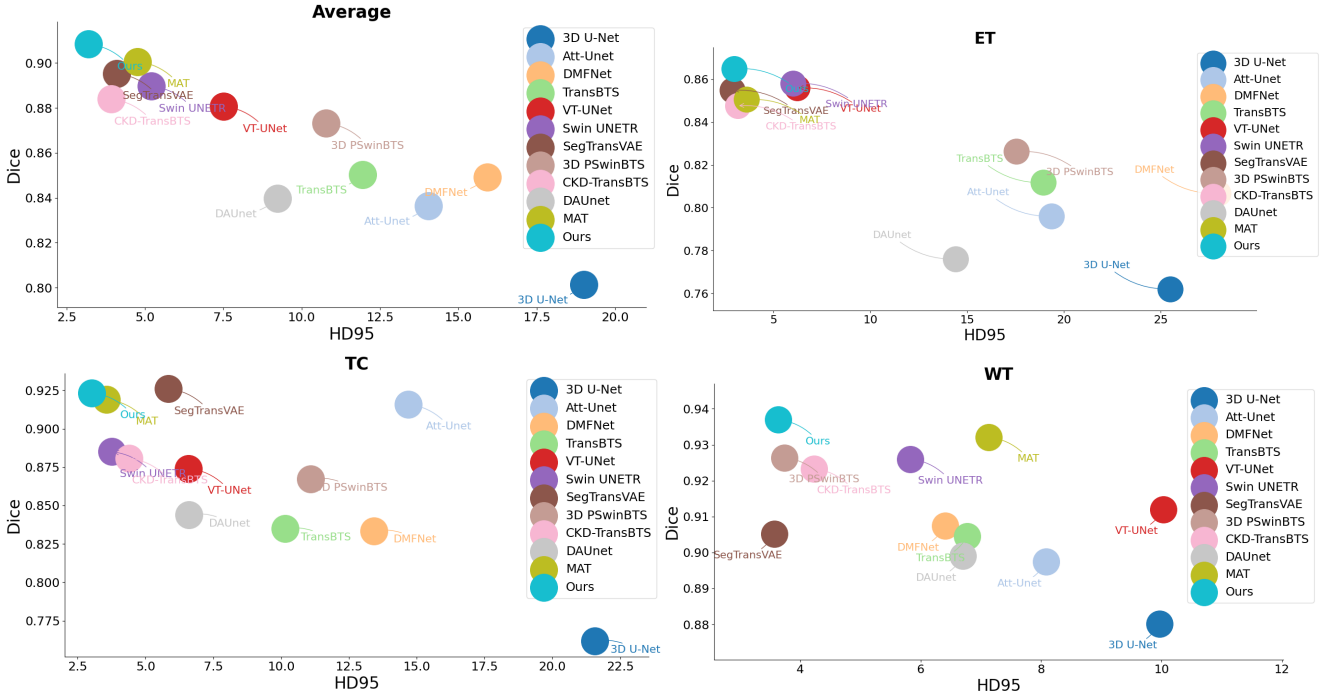

Fig. 7. Scatter plots with regard to the performance of different brain tumor segmentation methods on the BraTS2021 benchmark.

Based on the results reported in the above Tables IV and V and Figs. 6 and 7, our proposed method has demonstrated a significantly competitive performance compared to other approaches. Specifically, quantitative results indicate that our TransSea network achieved the best segmentation performance on the BraTS 2020-2021 benchmarks, with Dice scores of 86.32% and 90.84%, surpassing other reference methods by margins ranging from 0.16% to 7.67% and 3.52% to 10.71%, respectively. In addition, the average HD95 metric for the proposed method was reduced to 4.892 on BraTS 2020 and 3.207 on BraTS 2021 benchmarks. These comparative results reveal that hybrid networks combining Transformer and CNN architectures outperform both CNN- and Transformer-based networks, highlighting the efficacy and significance of integrating Transformers and CNNs for medical image segmentation. In other words, the model's ability to fuse global and local features contributes to enhanced segmentation results. When compared to Swin UNETR, which combines hierarchical Swin Transformer and convolution for semantic segmentation, and VT-UNet, which utilizes volume Transformers for encoding and decoding, our proposed method consistently outperforms them in all scenarios. On the BraTS 2021 benchmark, the average Dice scores improved from 88.97% and 88.07% to 90.84%, and the Hausdorff distance (HD) reduced from 5.206 and 7.520 mm to 3.207 mm, showcasing a substantial advantage. The most notable improvement was observed in the

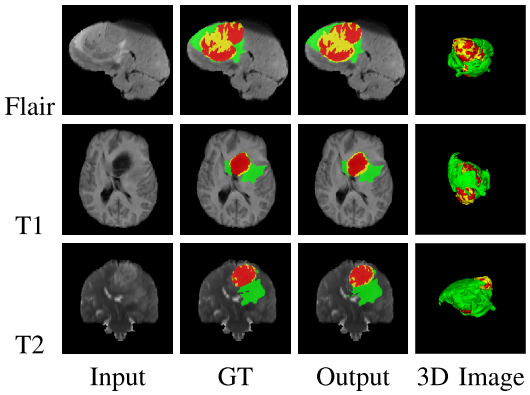

Fig. 8. 2-D and 3-D segmentation effects of different modality imaging of 3-D segmentation network. Green, yellow, and red indicate the ED, ET, and NCR/NET regions, respectively.

TC region, where it also demonstrates superior segmentation results in the WT region. This can be attributed to the convolution module of TransSea, which excels in both depth and spatial feature extraction capabilities, and its utilization of semantic information to guide feature fusion in the decoding section, making the segmentation more attuned to the global context. Furthermore, compared to the recent 3-D PSwinBTS method that incorporates semantic supervised learning, our proposed approach exhibits significant improvements across all aspects, underscoring the effectiveness of our strategy in directing the encoder's attention toward regions relevant to the segmentation task.

When employing 2-D methods for the segmentation of 3-D MRI images of brain tumors, there is a risk of losing crucial volumetric information by breaking down the 3-D image into 2-D slices and applying segmentation algorithms. In contrast, leveraging 3-D brain tumor segmentation techniques allows for a more effective utilization of spatial and volumetric information within the 3-D data, a fact supported by an analysis of Dice coefficients and HD95 performance metrics for 3-D networks. In the proposed model, the holistic processing of 3-D volumetric data and the extraction of 3-D MRI voxel features are paramount for performance. Therefore, enhancing the network architecture to bolster its capacity for handling spatial and volumetric information is of utmost importance. Fig. 8 visually contrasts our method's brain tumor segmentation efficacy in both 2-D and 3-D perspectives with real MRI images. These original images are sourced from the BraTS 2021 training dataset. Our method demonstrates accuracy and robustness in the domain of brain tumor segmentation.

Fig. 9 presents a visual comparison of the segmentation results obtained by different methods for brain tumor segmentation. By referencing the GT, our proposed method achieves more accurate segmentation results compared to other methods, particularly in the case of some inconspicuous abnormal regions that are overlooked by other segmentation approaches. This underscores the effectiveness of semantic context-aware feature extraction for segmentation.

C. Parameter Count and Computational Complexity

In Table VI, we summarize a quantitative analysis of the parameters and computational complexity of the experimental models mentioned above. Among all hybrid models based on CNN and Transformers, our proposed TransSea exhibits relatively smaller parameters and computational requirements, despite achieving high accuracy. Its parameter count and computational complexity are 9.9 M and 55.8 G FLOPs, respectively, lower than other comparative methods, indicating that our approach consumes fewer hardware resources and opens up possibilities for widespread clinical applications. However, compared to CNN-based methods, such approaches still face a disadvantage in terms of computational time, as convolutional operations are accelerated in PyTorch/TensorFlow. Combining the segmentation accuracy of the model with its computational complexity, our method achieves optimal segmentation performance with relatively moderate parameters and FLOPs.

TABLE VI
PARAMETER QUANTITY AND COMPUTATIONAL COMPLEXITY OF THE EXPERIMENTAL MODEL

Method	Parameter (M)	FLOPs (G)
3D U-Net [9]	2.4	162.7
Att-Unet [63]	6.4	151.2
DMFNet [71]	3.8	27.0
TransBTS [13]	30.6	263.8
VT-UNet [65]	11.8	100.8
3D PSwinBTS [48]	20.4	68.6
Proposed	9.9	55.8

TABLE VII
OBJECTIVE ASSESSMENT RESULTS OF THE PROPOSED METHOD FOR SEGMENTATION OF SOURCE IMAGES WITH GAUSSIAN NOISE (ZERO MEAN AND DIFFERENT STANDARD DEVIATION) ON THE BRATS2020 DATASET

	Dice (%) ↑				Hausdorff 95% (mm) ↓			
	Avg.	ET	TC	WT	Avg.	ET	TC	WT
w/o Noise	86.32	79.71	86.89	92.36	4.892	5.896	4.789	3.992
$\sigma = 10$	85.10	78.24	85.69	91.38	5.551	6.568	5.749	4.337
$\sigma = 20$	84.41	77.59	85.49	90.14	6.710	6.582	7.133	6.416

D. Robustness to Noise

Given that noise contamination is a prevalent issue in medical imaging, we conducted a series of experiments to explore the robustness of our method in terms of noise. For simplicity, we employed the widely used additive white Gaussian noise model. We assessed the performance of the proposed method when the source MRI images were contaminated with Gaussian noise at zero mean and different standard deviations ($\sigma = 5$, $\sigma = 10$, and $\sigma = 20$). As shown in Tables VII and VIII, a general performance degradation is observed with increasing levels of noise. Notably, our method demonstrates a consistently superior performance, maintaining higher Dice scores and lower Hausdorff distances even as noise intensity grows, demonstrating its superior noise resilience.

E. Ablation Study

In our comprehensive ablation study, we aim to validate and evaluate the proposed model's efficacy, focusing on how each component influences segmentation performance. We begin

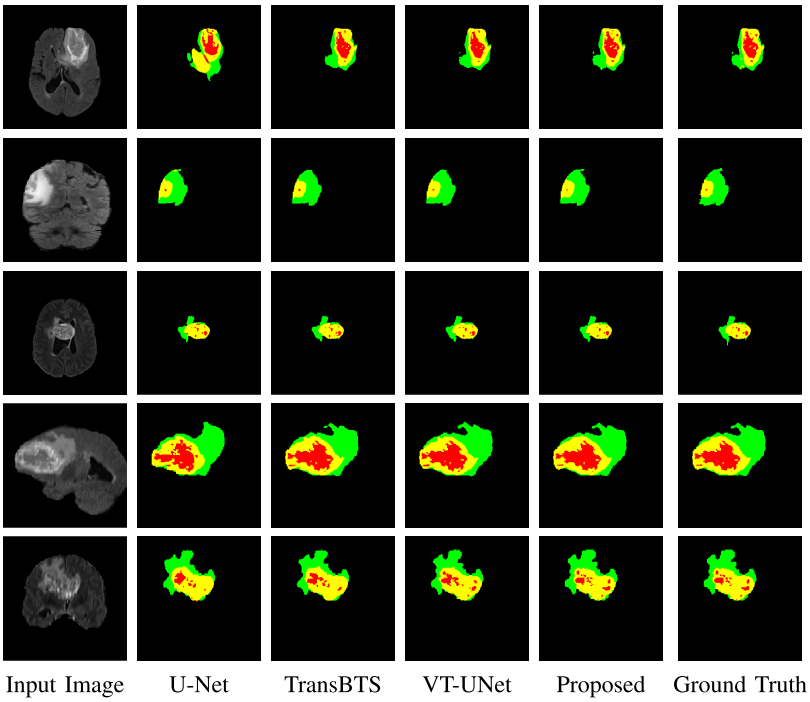

Fig. 9. Comparison of visual results of brain tumor segmentation results obtained by different methods.

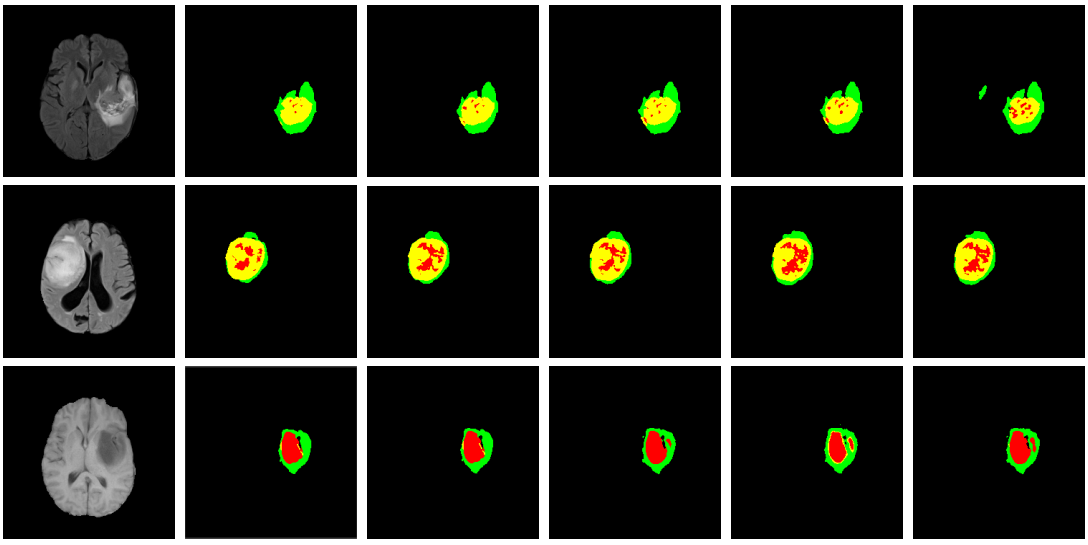

Fig. 10. Visual comparison of the segmentation results from different models in the ablation study, highlighting the 2-D and 3-D effects with the inclusion of various components.

TABLE VIII

OBJECTIVE ASSESSMENT RESULTS OF THE PROPOSED METHOD FOR SEGMENTATION OF SOURCE IMAGES WITH GAUSSIAN NOISE (ZERO MEAN AND DIFFERENT STANDARD DEVIATION) ON THE BRATS2021 DATASET

	Dice (%) ↑				Hausdorff 95% (mm) ↓			
	Avg.	ET	TC	WT	Avg.	ET	TC	WT
w/o Noise	90.84	86.49	92.33	93.70	3.207	2.964	3.024	3.632
$\sigma = 10$	90.18	86.04	91.28	93.23	3.350	3.071	3.200	3.779
$\sigma = 20$	88.36	84.51	87.81	92.75	4.608	3.730	4.623	5.470

with the standard Swin Transformer as our foundational baseline model. Gradually, we incorporate various module combinations, examining their contributions to enhancing performance. Key components under scrutiny include the SMA module, the SG module, and the SI module. A particular focus is given to the SMA module, where we experiment with its replacement using convolution blocks and standard Swin Transformer blocks, assessing their relative merits. The effects of these components are meticulously compared, employing both quantitative and qualitative lenses. To robustly test our model and its components, we deploy our suite of enhancements on the BraTS2021 dataset. The quantitative outcomes, detailed in Table IX, reveal the components' performance enhancement through Dice and HD95 metrics. A qualitative exposition of these findings is illustrated in Fig. 10. We specifically scrutinize the following model variations.

TABLE IX

OBJECTIVE EVALUATION RESULTS OF THE BRATS2021 BASELINE ABLATION STUDY. THIS TABLE PRESENTS THE QUANTITATIVE OUTCOMES FROM OUR ABLATION STUDY, WHICH EVALUATES THE IMPACT OF INCORPORATING INDIVIDUAL COMPONENTS OF OUR PROPOSED TRANSSEA NETWORK SMA, SG, AND SI INTO THE BASELINE MODEL. PERFORMANCE METRICS, SUCH AS DICE COEFFICIENT % AND HAUSDORFF 95% DISTANCE (mm), ARE PROVIDED FOR THREE TUMOR REGIONS: ET, TC, AND WT

	Dice (%) ↑				Hausdorff 95% (mm) ↓			
	Avg.	ET	TC	WT	Avg.	ET	TC	WT
Baseline	87.92	84.38	88.20	91.17	4.236	3.520	4.446	4.741
Baseline+SMA	89.12	85.24	89.91	92.22	3.986	3.486	3.826	4.645
Baseline+SMA+SG	90.23	85.62	91.90	93.15	3.342	3.044	3.133	3.850
Baseline+SMA+SG+SI	**90.84**	**86.49**	**92.33**	**93.70**	**3.207**	**2.964**	**3.024**	**3.632**

1) *Baseline Model:* Utilizing Swin Transformer blocks in an encoder–decoder configuration, serving as the foundational model for 3-D brain tumor segmentation.
2) *Baseline + SMA:* Enhanced with an SMA module, this variant merges global and local feature strategies but omits the SG branch in segmentation.
3) *Baseline + SMA + SG:* Building on the prior model, this version incorporates the SG module, mirroring our proposed model's framework, albeit with a simplified fusion of semantic and decoder features in the decoding phase, bypassing our advanced SI technique.
4) *Baseline + SMA + SG + SI:* The complete model we proposed.

1) Effectiveness of the SMA Module: Table IX delineates the objective performance metrics across different models. A careful examination reveals that each component, particularly the SMA module and SI-based feature fusion, significantly bolsters segmentation outcomes. We evaluate the SMA's impact by contrasting the baseline model against its SMA-enhanced counterpart. As detailed in Table IX, the inclusion of SMA leads to notable Dice score enhancements for ET, TC, and WT on the BraTS2021 dataset, recording increases of 0.86%, 1.71%, and 1.05%, respectively. This improvement is underscored by the quantitative comparison between the first and second rows, underscoring SMA's efficacy in global–local interaction modeling and semantic context feature extraction, validating that SMA can narrow the semantic gap between local and global features. In addition, aligning these findings with the qualitative insights from Fig. 10, it is evident that global semantic segmentation becomes more precise, especially in delineating boundaries of remote independent voxel points, compared to the baseline models initial column.

In our effort to evaluate the SMA module's effectiveness, we juxtapose the final model with variants where the SMA is replaced by convolution and Transformer blocks. Table X presents a detailed account of the performance of different feature extraction modules in the encoder. Notably, the comprehensive SMA module, incorporating convolution blocks, Transformer blocks, and CSA, demonstrates superior results. In addition, by removing CSA from the SMA and employing a simple concatenation strategy, we affirm the capability of the semantic interaction attention mechanism in bridging the semantic divide between local and global features. As shown in Table X, models excluding CSA struggle more with segmenting the TC. Moreover, we found that models with CSA exhibit a more stable training behavior compared to their CSA-excluded counterparts, highlighting the significance of a balanced integration of local and global features to enhance segmentation accuracy.

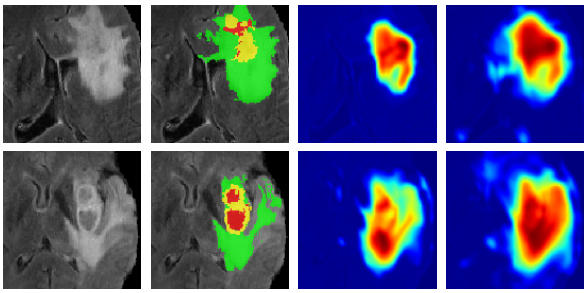

Fig. 11. Comparison of feature heat map before and after using SG.

TABLE X

ABLATION STUDY OF DIFFERENT FEATURE EXTRACTION MODULE IN SMA. THIS TABLE DETAILS THE EFFECTS OF DIFFERENT CONFIGURATIONS WITHIN THE SMA MODULE ON SEGMENTATION PERFORMANCE, FOCUSING ON VARIOUS ARCHITECTURAL VARIATIONS: WITHOUT CONVOLUTION (W/O CONV), WITHOUT TRANSFORMER (W/O TRANS), AND WITHOUT CSA (W/O CSA)

Module	Dice (%) ↑				Hausdorff 95% (mm) ↓			
	Avg.	ET	TC	WT	Avg.	ET	TC	WT
w/o Conv	89.88	85.88	90.73	93.05	4.380	4.644	3.638	4.859
w/o Trans	89.65	84.32	91.38	93.24	3.806	3.662	3.274	4.483
w/o CSA	89.46	85.24	89.91	93.22	4.094	3.486	3.826	4.970
SMA	**90.84**	**86.49**	**92.33**	**93.70**	**3.207**	**2.964**	**3.024**	**3.632**

2) Effectiveness of the SG Module: SG has been strategically incorporated to enhance early semantic supervision learning while also generating query values pivotal for guiding the decoding phase's feature fusion. A critical assessment of the experimental outcomes validates the SG module's utility. It is essential to note that the SG module is not standalone but requires integration with a semantic segmentation network. In terms of quantitative achievements on the BraTS2021 dataset, the integration of SG resulted in a substantial increase in the Dice and HD metrics for TC by 1.99% and 0.693%, respectively. Qualitative analysis, particularly evident in Fig. 11, contrasts the encoder feature heatmaps pre- and post-integration of SG. The model with SG distinctively intensifies focus on the pertinent features of the segmentation

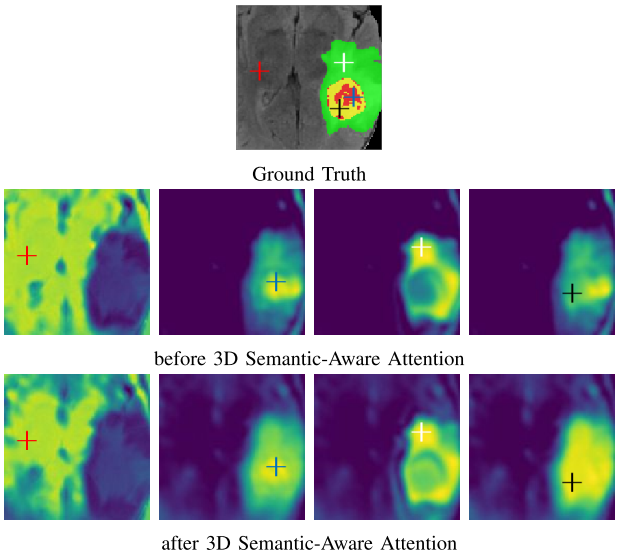

Fig. 12. Pixel-wise attention maps of the features corresponding to the cross-sign before/after 3-D semantic-aware attention. From left to right, the categories of pixels denoted by cross-signs in red, blue, white, and black represent the background, necrotic tumor, peritumoral edema, and ET, respectively.

target, an enhancement not observed in the model devoid of SG.

3) Effectiveness of the SI Module: Integrating the SI module results in notable advancements in segmentation outcomes. Specifically, on the BraTS2021 dataset, the Dice scores for TC and WT see a 1.03% enhancement, while ET benefits from a 0.87% and 0.55% improvement in Dice and HD metrics, respectively. The qualitative aspect, illustrated in Fig. 10, showcases the complete model's superior segmentation quality over its SI-absent counterpart, characterized by clearer, less disrupted boundaries closely aligning with the GT. This underscores the significance of query-based semantic-aware feature concentration for intricate tasks such as segmentation, stemming from the effective exploitation of semantic information, as previously elaborated. The model can focus more accurately on the regions to be segmented and mask out irrelevant semantic information. As shown in Fig. 12, we analyze pixel-level attention maps for four pixels with different classes before and after the semantic-aware attention map. It is observed that when features pass through the semantic-aware attention map, the current feature exhibits higher similarity with other features of the same class, resulting in finer attention areas.

V. Conclusion

This article proposes a novel semantic-aware brain tumor segmentation method, namely, TransSea, that harnesses the advantages of both CNNs and Transformers. Our unique SMA module sets itself apart from other hybrid Transformer networks by achieving parallel processing of Swin Transformer and DSConv. This design not only enhances the efficiency of merging global and local features but also ensures that each feature type is optimally utilized, marking a significant departure from traditional hybrid models. In addition, we have incorporated the SG module, which effectively utilizes semantic priors to guide semantic modeling, generate a semantic-guided attention map, and facilitate feature fusion during the decoding process through query values. Ultimately, our developed SI module integrates query-based semantic-guided features and decoder features through multiscale and cascading methods, focusing on semantic-aware features. The experimental results validate the effectiveness of the key components designed in our method. Moreover, the proposed method outperforms state-of-the-art methods on the BraTS benchmark. In summary, TransSea not only proposes a novel architecture but also sets a new benchmark for integrating semantic depth with technical sophistication in brain tumor segmentation, establishing a new direction for future research in medical image analysis. Segmenting multiple tissue types in brain MRI images is crucial not only for comprehensive brain analysis but also for enhancing the assessment of various neurological conditions. Given the promising results achieved in tumor segmentation, we are motivated to explore adapting our models for the segmentation of other brain tissues in the future work.

References

[1] Y. Tian et al., "Survey on deep learning in multimodal medical imaging for cancer detection," *Neural Comput. Appl.*, early access, pp. 1–16, Nov. 2023. [Online]. Available: https://link.springer.com/article/10.1007/s00521-023-09214-4, doi: 10.1007/s00521-023-09214-4.

[2] F. Piccialli, V. D. Somma, F. Giampaolo, S. Cuomo, and G. Fortino, "A survey on deep learning in medicine: Why, how and when?" *Inf. Fusion*, vol. 66, pp. 111–137, Feb. 2021.

[3] S. Mo et al., "Multimodal priors guided segmentation of liver lesions in MRI using mutual information based graph co-attention networks," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Lima, Peru. Cham, Switzerland: Springer, Oct. 2020, pp. 429–438.

[4] A. Myronenko, "3D MRI brain tumor segmentation using autoencoder regularization," in *Proc. Int. MICCAI Brainlesion Workshop*, 2018, pp. 311–320.

[5] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, vol. 9351. Cham, Switzerland: Springer, 2015, pp. 234–241.

[6] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[7] Z. Zhou, M. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-net architecture for medical image segmentation," in *Proc. Int. Workshop Deep Learn. Med. Image Anal.*, vol. 11045, 2018, pp. 3–11.

[8] O. Oktay et al., "Attention U-Net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999*.

[9] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2016, pp. 424–432.

[10] F. Milletari, N. Navab, and S. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 565–571.

[11] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[12] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[13] W. Wang, C. Chen, M. Ding, H. Yu, S. Zha, and J. Li, "TransBTS: Multimodal brain tumor segmentation using transformer," in *Proc. 24th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2021, pp. 109–119.

[14] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

[15] S. H. Lee, S. Lee, and B. C. Song, "Vision transformer for small-size datasets," 2021, *arXiv:2112.13492*.

[16] H. Wu et al., "CvT: Introducing convolutions to vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 22–31.

[17] W. Wang et al., "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 568–578.

[18] J. Chen et al., "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.

[19] S. Li, X. Sui, X. Luo, X. Xu, Y. Liu, and R. Goh, "Medical image segmentation using squeeze-and-expansion transformers," in *Proc. 13th Int. Joint Conf. Artif. Intell.*, Aug. 2021, pp. 807–815.

[20] J. Dolz, K. Gopinath, J. Yuan, H. Lombaert, C. Desrosiers, and I. B. Ayed, "HyperDense-Net: A hyper-densely connected CNN for multi-modal image segmentation," *IEEE Trans. Med. Imag.*, vol. 38, no. 5, pp. 1116–1126, May 2019.

[21] E. Xie et al., "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. NIPS*, 2021, pp. 12077–12090.

[22] M. Fan, W. Wang, W. Yang, and J. Liu, "Integrating semantic segmentation and retinex model for low-light image enhancement," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2317–2325.

[23] S. Zheng and G. Gupta, "Semantic-guided zero-shot learning for low-light image/video enhancement," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. Workshops (WACVW)*, Waikoloa, HI, USA, Jan. 2022, pp. 581–590.

[24] F. Isensee, P. Kickingereder, W. Wick, M. Bendszus, and K. H. Maier-Hein, "No new-Net," in *Proc. Int. MICCAI Brainlesion Workshop*, Granada, Spain. Cham, Switzerland: Springer, Sep. 2018, pp. 234–244.

[25] W. Feifan, J. Runzhou, Z. Liqin, M. Chun, and B. Bharat, "3D U-Net based brain tumor segmentation and survival days prediction," in *Proc. Int. MICCAI Brainlesion Workshop*, 2019, pp. 131–141.

[26] Y. Liu, Y. Shi, F. Mu, J. Cheng, C. Li, and X. Chen, "Multimodal MRI volumetric data fusion with convolutional neural networks," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–15, 2022.

[27] K. Chang et al., "Automatic assessment of glioma burden: A deep learning algorithm for fully automated volumetric and bidimensional measurement," *Neuro-Oncol.*, vol. 21, no. 11, pp. 1412–1422, Nov. 2019.

[28] P. Kickingereder et al., "Automated quantitative tumour response assessment of MRI in neuro-oncology with artificial neural networks: A multicentre, retrospective study," *Lancet Oncol.*, vol. 20, no. 5, pp. 728–740, May 2019.

[29] J. Shan, H. D. Cheng, and Y. Wang, "Completely automated segmentation approach for breast ultrasound images using multiple-domain features," *Ultrasound Med. Biol.*, vol. 38, no. 2, pp. 262–275, Feb. 2012.

[30] K. R. Singh, A. Sharma, and G. K. Singh, "MADRU-Net: Multiscale attention-based cardiac MRI segmentation using deep residual U-Net," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–13, 2024.

[31] H. Yin and Y. Shao, "CFU-Net: A coarse–fine U-Net with multilevel attention for medical image segmentation," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–12, 2023.

[32] Z. Zhou et al., "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, Dec. 2020.

[33] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual U-Net," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, pp. 749–753, 2018.

[34] S. Metlek and H. Çetiner, "ResUNet+: A new convolutional and attention block-based approach for brain tumor segmentation," *IEEE Access*, vol. 11, pp. 69884–69902, 2023.

[35] S. Montaha, S. Azam, A. K. M. R. H. Rafid, M. Z. Hasan, and A. Karim, "Brain tumor segmentation from 3D MRI scans using U-Net," *Social Netw. Comput. Sci.*, vol. 4, no. 4, p. 386, May 2023.

[36] R. Wang, S. Chen, C. Ji, J. Fan, and Y. Li, "Boundary-aware context neural network for medical image segmentation," *Med. Image Anal.*, vol. 78, May 2022, Art. no. 102395.

[37] Y. Liu, F. Mu, Y. Shi, J. Cheng, C. Li, and X. Chen, "Brain tumor segmentation in multimodal MRI via pixel-level and feature-level image fusion," *Frontiers Neurosci.*, vol. 16, Sep. 2022, Art. no. 1000587.

[38] I. Aboussaleh, J. Riffi, K. E. Fazazy, M. A. Mahraz, and H. Tairi, "Efficient U-Net architecture with multiple encoders and attention mechanism decoders for brain tumor segmentation," *Diagnostics*, vol. 13, no. 5, p. 872, Feb. 2023.

[39] X. Dong et al., "CSWin transformer: A general vision transformer backbone with cross-shaped windows," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12124–12134.

[40] 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 *Proc. Int. MICCAI Brainlesion Workshop*, 2021, pp. 272–284.

[41] X. Pan et al., "On the integration of self-attention and convolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 815–825.

[42] Z. Zhu, X. He, G. Qi, Y. Li, B. Cong, and Y. Liu, "Brain tumor segmentation based on the fusion of deep semantics and edge information in multimodal MRI," *Inf. Fusion*, vol. 91, pp. 376–387, Mar. 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253522001981

[43] J. Guo et al., "CMT: Convolutional neural networks meet vision transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12175–12185.

[44] H. Zhang et al., "Context encoding for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Mar. 2018, pp. 7151–7160.

[45] Z. Jin et al., "Mining contextual information beyond image for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7231–7241.

[46] Z. Jin, B. Liu, Q. Chu, and N. Yu, "ISNet: Integrate image-level and semantic-level context for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7189–7198.

[47] Y. Feng, Y. Cao, D. An, P. Liu, X. Liao, and B. Yu, "DAUnet: A U-shaped network combining deep supervision and attention for brain tumor segmentation," *Knowl.-Based Syst.*, vol. 285, Feb. 2024, Art. no. 111348. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705123010961

[48] J. Liang, C. Yang, and L. Zeng, "3D PSwinBTS: An efficient transformer-based UNet using 3D parallel shifted windows for brain tumor segmentation," *Digit. Signal Process.*, vol. 131, Nov. 2022, Art. no. 103784.

[49] J. Jain et al., "SeMask: Semantically masked transformers for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 752–761.

[50] H. Huang et al., "UNet 3+: A full-scale connected UNet for medical image segmentation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 1055–1059.

[51] P. Gao et al., "Dynamic fusion with intra- and inter-modality attention flow for visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6632–6641.

[52] X. Xu, T. Wang, Y. Yang, L. Zuo, F. Shen, and H. T. Shen, "Cross-modal attention with semantic consistence for image–text matching," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 12, pp. 5412–5425, Dec. 2020.

[53] J. Liang, C. Yang, M. Zeng, and X. Wang, "TransConver: Transformer and convolution parallel network for developing automatic brain tumor segmentation in MRI images," *Quant. Imag. Med. Surgery*, vol. 12, no. 4, pp. 2397–2415, Apr. 2022.

[54] Z. Liu et al., "Video Swin transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 3202–3211.

[55] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters—Improve semantic segmentation by global convolutional network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1743–1751.

[56] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[57] C. Yu, J. Wang, C. Gao, G. Yu, C. Shen, and N. Sang, "Context prior for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12416–12425.

[58] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Sep. 2017, pp. 1251–1258.

[59] C. Xiao, S. Yang, and Z. Feng, "Complex-valued depthwise separable convolutional neural network for automatic modulation classification," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–10, 2023.

[60] Y. Chen et al., "Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3435–3444.

[61] B. Cheng, R. Xiao, J. Wang, T. Huang, and L. Zhang, "High frequency residual learning for multi-scale image classification," 2019, *arXiv:1905.02649*.

[62] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5728–5739.

[63] S. Lian, Z. Luo, Z. Zhong, X. Lin, S. Su, and S. Li, "Attention guided U-Net for accurate iris segmentation," *J. Vis. Commun. Image Represent.*, vol. 56, pp. 296–304, Oct. 2018.

[64] Y. Jiang, Y. Zhang, X. Lin, J. Dong, T. Cheng, and J. Liang, "SwinBTS: A method for 3D multimodal brain tumor segmentation using Swin transformer," *Brain Sci.*, vol. 12, no. 6, p. 797, Jun. 2022.

[65] H. Peiris, M. Hayat, Z. Chen, G. Egan, and M. Harandi, "A robust volumetric transformer for accurate 3D tumor segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2022, pp. 162–172.
[66] Z. Xing, L. Yu, L. Wan, T. Han, and L. Zhu, "NestedFormer: Nested modality-aware transformer for brain tumor segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Singapore. Cham, Switzerland: Springer, Sep. 2022, pp. 140–150.
[67] Y. Zhuang, H. Liu, E. Song, and C.-C. Hung, "A 3D cross-modality feature interaction network with volumetric feature alignment for brain tumor and tissue segmentation," *IEEE J. Biomed. Health Informat.*, vol. 27, no. 1, pp. 75–86, 2023.
[68] H. Liu, G. Huo, Q. Li, X. Guan, and M.-L. Tseng, "Multiscale lightweight 3D segmentation algorithm with attention mechanism: Brain tumor image segmentation," *Expert Syst. Appl.*, vol. 214, Mar. 2023, Art. no. 119166.
[69] I. Aboussaleh, J. Riffi, K. el Fazazy, A. M. Mahraz, and H. Tairi, "3DUV-NetR+: A 3D hybrid semantic architecture using transformers for brain tumor segmentation with MultiModal MR images," *Results Eng.*, vol. 21, Mar. 2024, Art. no. 101892. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2590123024001452
[70] F. Isensee, P. F. Jäger, P. M. Full, P. Vollmuth, and K. H. Maier-Hein, "nnU-Net for brain tumor segmentation," in *Proc. Int. MICCAI Brainlesion Workshop*, Oct. 2020, pp. 118–132.
[71] C. Chen, L. Xiaopeng, D. Meng, Z. Junfeng, and L. Jiangyun, "3D dilated multi-fiber network for real-time brain tumor segmentation in MRI," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*, 2019, pp. 184–192.
[72] Y. Tang et al., "Self-supervised pre-training of Swin transformers for 3D medical image analysis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 20730–20740.
[73] Q.-D. Pham et al., "Segtransvae: Hybrid CNN–transformer with regularization for medical image segmentation," in *Proc. IEEE 19th Int. Symp. Biomed. Imag. (ISBI)*, Mar. 2022, pp. 1–5.
[74] J. Lin et al., "CKD-TransBTS: Clinical knowledge-driven hybrid transformer with modality-correlated cross-attention for brain tumor segmentation," *IEEE Trans. Med. Imag.*, vol. 42, no. 8, pp. 2451–2461, Aug. 2023.
[75] C. Liu and H. Kiryu, "3D medical axial transformer: A lightweight transformer model for 3D brain tumor segmentation," in *Proc. Med. Imag. With Deep Learn.*, 2024, pp. 799–813.
[76] U. Baid et al., "The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification," 2021, *arXiv:2107.02314*.
[77] F. Bieder, J. Wolleb, A. Durrer, R. Sandkühler, and P. C. Cattin, "Diffusion models for memory-efficient processing of 3D medical images," 2023, *arXiv:2303.15288*.
[78] H. Luo, D. Zhou, Y. Cheng, and S. Wang, "MPEDA-Net: A lightweight brain tumor segmentation network using multi-perspective extraction and dense attention," *Biomed. Signal Process. Control*, vol. 91, May 2024, Art. no. 106054.
[79] R. Raza, U. I. Bajwa, Y. Mehmood, M. W. Anwar, and M. H. Jamal, "DResU-Net: 3D deep residual U-Net based brain tumor segmentation from multimodal MRI," *Biomed. Signal Process. Control*, vol. 79, Jan. 2023, Art. no. 103861.

Yu Liu (Member, IEEE) received the B.S. and Ph.D. degrees from the Department of Automation, University of Science and Technology of China, Hefei, China, in 2011 and 2016, respectively.

He is currently an Associate Professor with the Department of Biomedical Engineering, Hefei University of Technology, Hefei. His research interests include image processing, computer vision, information fusion, and machine learning. In particular, he is interested in image fusion, image restoration, visual recognition, and deep learning.

Dr. Liu was a recipient of the IEEE Instrumentation and Measurement Society Andi Chi Best Paper Award in 2020 and the *IET Image Processing* Premium (Best Paper) Award in 2017. He was identified as a Clarivate Highly Cited Researcher in 2023. He is serving as an Editorial Board Member for *Information Fusion* and an Associate Editor for IEEE SIGNAL PROCESSING LETTERS.

Yize Ma received the B.S. degree from Jiangsu University, Zhenjiang, China, in 2022. He is currently pursuing the M.S. degree in biomedical engineering with Hefei University of Technology, Hefei, China.

His current research interests include medical image segmentation and deep learning.

Zhiqin Zhu received the B.E. and Ph.D. degrees from Chongqing University, Chongqing, China, in 2010 and 2016, respectively.

He is currently a Full Professor with the Automation College, Chongqing University of Post and Telecommunications, Chongqing. His primary research interests include machine learning, image processing, and medical imaging.

Dr. Zhu was a recipient of the IEEE Instrumentation and Measurement Society Andi Chi Best Paper Award in 2022.

Juan Cheng (Member, IEEE) received the B.S. and Ph.D. degrees from the Department of Electronic Science and Technology, University of Science and Technology of China, Hefei, China, in 2008 and 2013, respectively.

She is currently a Professor with the Department of Biomedical Engineering, Hefei University of Technology, Hefei. Her research interests include biomedical signal/image processing, mobile healthcare, and machine learning.

Xun Chen (Senior Member, IEEE) received the B.Sc. degree in electrical engineering from the University of Science and Technology of China, Hefei, China, in 2009, and the Ph.D. degree in biomedical engineering from The University of British Columbia (UBC), Vancouver, BC, Canada, in 2014.

He has been a Research Scientist with the Department of Electrical and Computer Engineering, UBC. He is currently a Full Professor and the Head of the Department of Electrical Engineering and Information Science, University of Science and Technology of China. He has published over 100 scientific articles in prestigious IEEE/Elsevier journals and conferences. His research interests include the broad areas of statistical signal processing and machine learning in biomedical applications.

Dr. Chen is serving as an Area Editor for *Signal Processing: Image Communication* and an Associate Editor for IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENT, IEEE SIGNAL PROCESSING LETTERS, and *Frontiers in Neuroscience*.