IEEE Journal of Translational Engineering in Health and Medicine

Received 7 November 2022; revised 15 April 2023 and 31 May 2023; accepted 17 June 2023.
Date of publication 27 June 2023; date of current version 18 July 2023.

Digital Object Identifier 10.1109/JTEHM.2023.3289990

TransU^2-Net: An Effective Medical Image Segmentation Framework Based on Transformer and U^2-Net

XIANG LI[1], XIANJIN FANG[2,3], GAOMING YANG[2], SHUZHI SU[2], LI ZHU[4], AND ZEKUAN YU[2,5]

[1]School of Safety Science and Engineering, Anhui University of Science and Technology, Huainan 232000, China
[2]School of Computer Science and Engineering, Anhui University of Science and Technology, Huainan 232000, China
[3]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei 230009, China
[4]Shanghai Chest Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai 200030, China
[5]Academy for Engineering and Technology, Fudan University, Shanghai 200433, China

CORRESPONDING AUTHORS: X. FANG (xjfang@aust.edu.cn), L. ZHU (augjuly@aliyun.com), AND Z. YU (yzk@fudan.edu.cn)

This work was supported in part by the University Synergy Innovation Program of Anhui Province under Grant GXXT-2021-006, in part by the Institute of Engineering and Technology of Shanghai Fudan University, in part by the Shanghai Hospital Development Centre under Grant SHDC2020CR3020A, and in part by the Joint Fund for Medical Engineering of Fudan University.

ABSTRACT Background: In the past few years, U-Net based U-shaped architecture and skip-connections have made incredible progress in the field of medical image segmentation. U^2-Net achieves good performance in computer vision. However, in the medical image segmentation task, U^2-Net with over nesting is easy to overfit. Purpose: A 2D network structure TransU^2-Net combining transformer and a lighter weight U^2-Net is proposed for automatic segmentation of brain tumor magnetic resonance image (MRI). Methods: The light-weight U^2-Net architecture not only obtains multi-scale information but also reduces redundant feature extraction. Meanwhile, the transformer block embedded in the stacked convolutional layer obtains more global information; the transformer with skip-connection enhances spatial domain information representation. A new multi-scale feature map fusion strategy as a postprocessing method was proposed for better fusing high and low-dimensional spatial information. Results: Our proposed model TransU^2-Net achieves better segmentation results, on the BraTS2021 dataset, our method achieves an average dice coefficient of 88.17%; Evaluation on the publicly available MSD dataset, we perform tumor evaluation, we achieve a dice coefficient of 74.69%; in addition to comparing the TransU^2-Net results are compared with previously proposed 2D segmentation methods. Conclusions: We propose an automatic medical image segmentation method combining transformers and U^2-Net, which has good performance and is of clinical importance. The experimental results show that the proposed method outperforms other 2D medical image segmentation methods.

Clinical Translation Statement: We use the BarTS2021 dataset and the MSD dataset which are publicly available databases. All experiments in this paper are in accordance with medical ethics.

INDEX TERMS Deep learning, medical image segmentation, transformer, U-Net.

I. INTRODUCTION

The majority of primary brain tumors, comprising between 30% and 40% of all brain tumors and over 80% of all malignant brain tumors in adulthood, are gliomas, and surgical resection is currently the most effective treatment, so automatic and accurate segmentation of brain tumors is very important in clinical evaluation and diagnosis [1], [2]. A typical neuroimaging method used in clinical practice for quantitative evaluation of common brain tumors is magnetic resonance imaging (MRI), which not only has the advantages of non-invasive, non-electrical radiation and high soft tissue contrast, but also provides a variety of different imaging models, such as T1-weighted (T1), contrast-enhanced T1-weighted (T1c), T2- weighted (T2), and Fluid attenuation inversion recovery (Flair), the different imaging modalities all provide the physician with different critical

information to obtain the most accurate diagnosis [3]. Therefore, for clinical applications, automated and precise segmentation of malignant tumors on multimodal MRI is crucial.

Preserving local features while conducting efficient segmentation is a crucial issue for medical image segmentation tasks [4]. In recent years, medical image segmentation models have been essential in advancing deep learning research. Particularly convolutional neural networks have been used in medically assisted diagnosis, and therapy as deep convolutional neural networks can obtain more features at different levels from the data [5]. Fully Convolutional Networks (FCN) [6] achieve end-to-end semantic segmentation; dense pixel-level prediction of medical pictures is made possible by U-Net [7], which uses a symmetric encoder-decoder structure with skip-connections to gradually restore the downsample feature maps to their initial dimensions [8]. U-Net is used as the basis for subsequent studies; encoder decoder-based U-Net achieves the acquisition of many U-Net-based 2/3D variant networks such as D-SEAU-Net [9], U-Net++ [10], ERU-Net [11], 3DV-Net [12], 3DU-Net [13], etc. The segmentation of brain tumors [14], liver [15], brain tissue [16], retinal blood vessel extraction [17], etc., has achieved high-quality performance.

However, the perceptual field of convolution is constrained, limiting its ability to capture long-distance relationships between different visual regions. Recently, the Vision Transformer (ViT) [18] has emerged as a solution, excelling at capturing global information by dynamically calculating weights between global pixels [19], [22], [26]. Compared to existing convolutional architectures, Vision Transformer achieves superior performance by partitioning images into predetermined-sized blocks and employing self-attention methods to establish relationships among these patches, surpassing the capabilities of conventional convolutional architectures [23], [24], [25]. This advancement has spurred the development of numerous transformer-based networks, including TransFuse [20], TransBTS [21], and Cotr [27], which combine transformers and CNNs to extract features.

Nonetheless, actual clinical medical images typically possess few scanning layers, low resolution, and disjointed contexts. In 3D convolutional-based networks, the improvement in accuracy is relatively marginal compared to the numerous models, and the convergence performance of transformer-based combinations on small datasets is poor, impeding the clinical application of transformers. Our research focuses on effectively integrating transformers with deep convolutional layers to achieve precise segmentation in clinical medical images.

An excellent segmentation model must be able to incorporate multi-scale features with fine-grained local details at the same time. Previously, feature fusion was accomplished by simply concatenating features, but FPN [28] creates a new feature pyramid approach for multi-feature fusion by extracting features of varying scales from different layers of the network architecture for prediction. Furthermore, FastFCN [29] is more semantic by combining features of different sizes after convolution.

In actuality, high-level and low-level feature information are complemented, which is crucial to effectively integrate both for semantic segmentation. Since low-level features resolution is higher, it contains more location and detailed information. But due to less convolution, it has lower semantics and more noise. High-level characteristics provide more semantic information but have limited resolution and poor detail perception [8], [30].

To address the above issues, we propose a new image segmentation method that demonstrates the feasibility of transformer embedded in deep convolution. This paper main contributions are as follows:

- We propose a novel and efficient segmentation network named TransU^2-Net. The new method can accurately capture the spatial domain information which is enhanced by introducing transformers across the skip connections;
- We construct a new multi scale feature map fusion strategy for enhancing fusing high and low dimensional spatial information, so that the output results have stronger low level feature semantic information and high-level details information;
- TransU^2-Net is the first transformer application in deeply nested U-shaped structures. The methods proposed in this paper have higher accuracy than most 3D models in many clinical medical segmentation applications.

II. MATERIALS AND METHOD

A. DATASET

1) BraTS2021 DATASET

We used the publicly available BraTS2021dataset [31], [32], [33], which offers a substantial amount of labelled brain tumor MRI datasets, primarily from cancer imaging archives. Since the BraTS2021 challenge's validation and test sets are unavailable to the general public, we subdivided all of the challenge's training sets (1251 3D MRI images), split into 7:3 ratios (876 cases for the training data and 375 cases for the test data). The four MRI modalities used for each patient's scans were T2 fluid-attenuated inversion recovery (T2-Flair), T1-weighted (T1), post-contrast T1-weighted (T1-Gd) and T2-weighted (T2), with T1 serving as the alignment standard. All data were resampled to 1 mm3 resolution after cranial debulking, with a final image size of 240 $\times$ 240 $\times$ 155; the ground-truths included four regions: enhancing tumor (ET-label 4), the necrotic tumor core (NCR-label 1), the healthy tissue, the peritumoral edematous and invading tissue (ED-label 2), and the background (Background-label 0). We segment and evaluate the whole tumor (label 2+4+1), tumor core (label 4+1), and enhancing tumor (label 1) segmentation results.

IEEE Journal of Translational Engineering in Health and Medicine

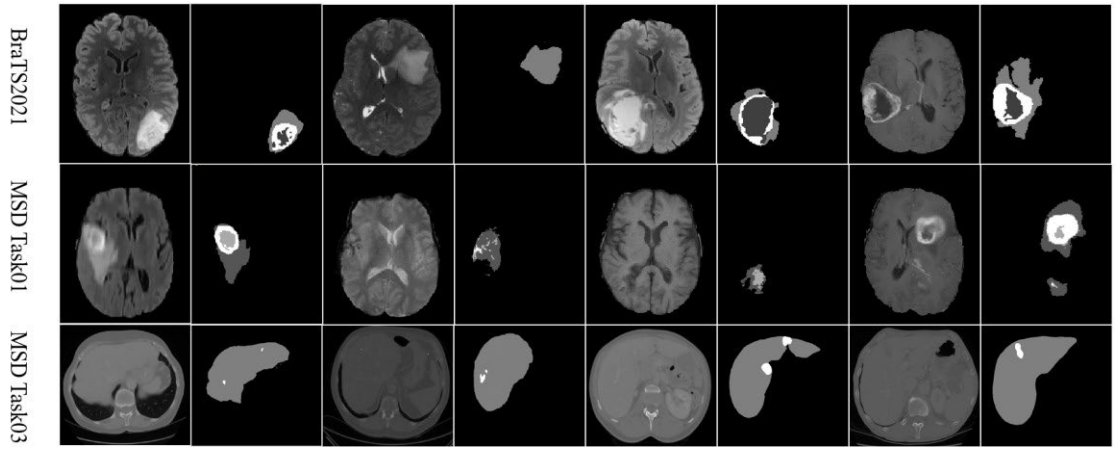

FIGURE 1. Thumbnails of medical images in various datasets.

2) MEDICAL SEGMENTATION DECATHLON DATASET

We used Task01_BrainTumour data and Task03_Liver data in MSD dataset [34]. Task01_BrainTumour provides a large number of patients diagnosed with either glioblastoma or lower grade glioblastoma, Task03_Liver provides patients with liver and liver tumors. Due to the Medical Segmentation Decathlon Dataset the verification and test sets are not publicly available, we segmented all the training sets (Task01_BrainTumour: 485 4D MRI images, Task03_Liver: 131 3D CT images) provided in the challenge and divided them according to 7:3 (Task01_BrainTumour: 340 training data and 144 testing data, Task03_Liver: 100 training data and 31 testing data). Task01_BrainTumour, all experimental cases were 4D MRI images, and using the SRI24 brain structure template, MRI scans were co-registered to a reference atlas space, resampled to isotropic voxel resolution of 1 mm^3, and skull-stripped using identical technique before being manually refined. Ground-truths included three regions: glioma (label 2), necrotic/activate (label 3) and edema (label 1). Segmentation accuracy was measured of whole tumor (label 1+2+3), and tumor core (label 2+3). Task03_Liver, the in-plane resolution of all 3D images is 0.5 to 1.0 mm, and the slice thickness is 0.45 to 6.0 mm. Liver and tumor annotations were performed by radiologists. Ground-truths included two regions: liver (label 1) and tumor (label 2). Segmentation accuracy was measured of whole liver (label 1+2) and tumor (label 2).

Some thumbnails of medical images and their ground truths from BraTS2021 Dataset and Medical Segmentation Decathlon Dataset are shown in Figure 1, respectively.

B. PREPROCESSING STEPS

To preprocess the 3D medical image data, we implemented 2D slicing. As the MSD Task01_BrainTumour data was in 4D, we split the data into 3D data for processing. For all 2D slices, we applied Z-Score normalization to each modality image. To reduce the proportion of background information in medical images, which occupies a relatively large portion of the overall image, we centrally cropped our images. Finally, we sliced the data from each modality and combined them into multiple channels.

C. SEGMENTATION NETWORK ARCHITECTURE

In this paper, we use a deeply nested U-shaped structure U^n-Net. In theory, n can take any positive integer, but too much-nested structure will consume many resources in the training process. According to the above, we set n=2, as shown in Figure 2, TransU^2-Net network is a two-level nested U-shaped structure. The input dimension of TransU^2-Net is $4 \times 160 \times 160$. The backbone network consists of seventeen blocks, which contain seventeen Conv Blocks and five Transformer Blocks. To obtain more global information, we use transformer in the U-Net cascade to get global information. In addition, we accept feature map information at different levels by the jump fusion strategy makes the low-level features fused with the high-level features so that the final segmentation result has stronger linguistic information and more detailed information at the same time. In the following subsections, we describe the specific implementation of each part in detail.

1) CONV MODULE

To handle complex medical image segmentation tasks, we use the Conv block as each layer of the encoder and decoder. The module is similar to a small U-Net structure; this allows us to extract more details, enabling us to extract finer details. To prevent overfitting, we incorporate dilated convolution in the middle layer of each conv block, which enhances the perceptual field. Each Conv Block comprises the structure illustrated in Figure 3, using a step size of 1 and padding of 1, each convolutional kernel is 3×3 in size. As the stage increases, we decrease the number of convolution layers per Conv Block, and at stages 5 and 6, we replace pooling and upsampling procedures with dilated convolutions. This indicates that all intermediate feature maps at stages 5 and

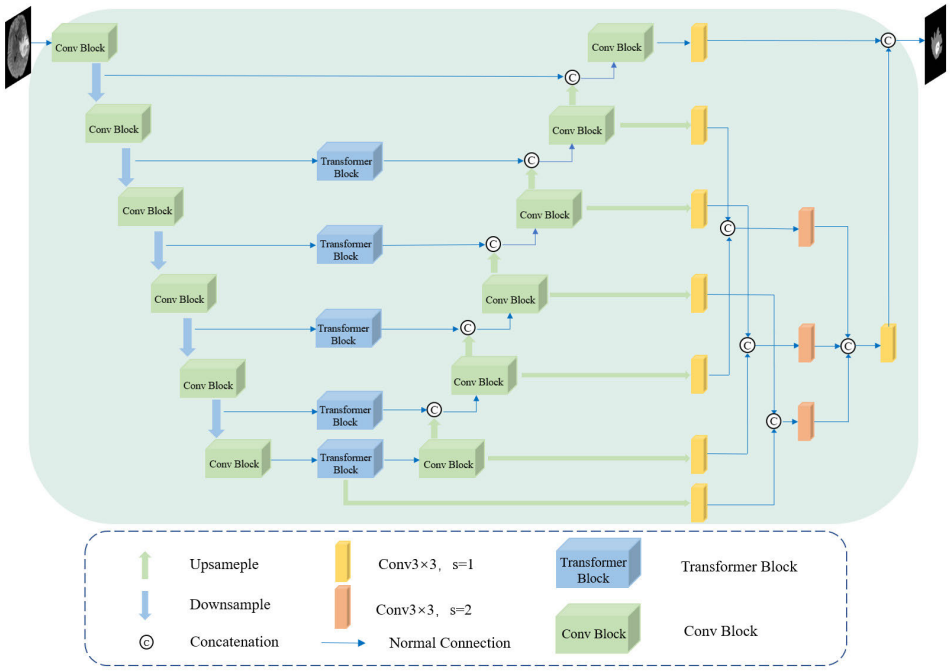

FIGURE 2. The illustration of the proposed TransU^2-Net for automatic medical image segmentation, we use U^2-Net to capture local information and leverage the Transformer encoder to model long-distance dependencies from the global view. Jump feature fusion module are stacked to gradually produce high-resolution segmentation results.

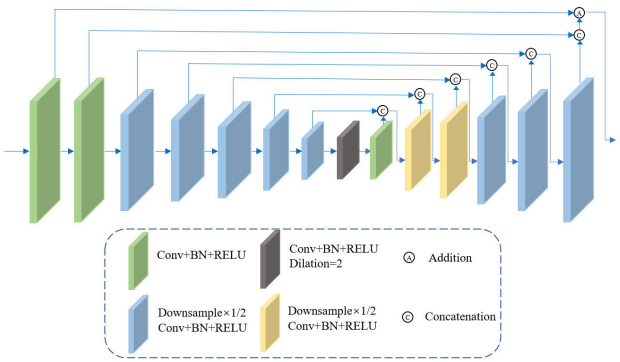

FIGURE 3. Conv module.

6 have the same resolution as their input feature maps. Specific parameters of each conv block are presented in Table 1.

2) TRANSFORMER MODULE

The original transformer model is only suitable for processing longer sequences of information. We incorporate positional encoding for each image patch, enabling the model to learn the relative positional relationships among image blocks, perception enhancement is then performed by the multi-layer perception mechanism. We have integrated a transformer module into the skip connection to capture global information in addition to shallow features. Figure 4 shows the specific implementation process of transformer block module, and

TABLE 1. Specific parameter configuration of conv block.

No.	Input size	Output size	CLN	Kernel size	Padding	Dilation
Stage1	160×160	80×80	6	3 × 3	1	1
Stage2	80×80	40×40	5	3 × 3	1	1
Stage3	40×40	20×20	4	3 × 3	1	1
Stage4	20×20	10×10	3	3 × 3	1	1
Stage5	10×10	5×5	2	3 × 3	2	2
Stage6	5×5	5×5	2	3 × 3	4	4

Stage represents the primary network layer count, Input size represents the tensor size of the input, Output size represents the tensor size of the output, CLN represents the number of layers of the convolution layer, Kernel represents the convolution kernel size, Padding and Dilation represent the parameters of the dilated convolution.

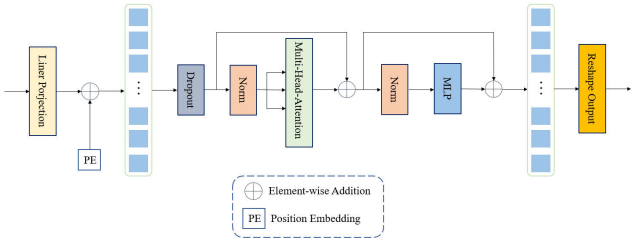

FIGURE 4. Transformer module.

the specific parameter configuration is shown in Table 2. Suppose the size of the token entering the Transformer Block is $\chi \epsilon \mathbb{R}^{H \times w \times S}$, the size of patches (H', W'), and the size of each token is $(\frac{H'}{H}, \frac{W'}{W})$, after which the token is reshapes,

IEEE Journal of Translational Engineering in Health and Medicine

TABLE 2. Specific parameter configuration of transformer block.

No.	Output size	Patch size	Trans dim	Heads	MLP dim
Side2	5×5	5	8	16	8
Side3	10×10	10	4	16	4
Side4	20×20	20	4	16	2
Side5	40×40	20	2	16	2
Side6	80×80	10	2	16	2

Side represents the number of connected layers. Input size represents the input tensor size. Output size represents the output tensor size. Patch size represents the patch size entering transformer. Trans_dim represents the number of transformer cycles. Heads represents the number of heads of multi-head self-attention. MLP dim represents the number.

and by Positional encoding we record the positional of each token, and in the subsequent Transformer Box the calculation process is as follows:

$$\hat{O}^l = MSA[Liner\_nor(O^l)] + O^{l-1} \tag{1}$$

$$\hat{O}^l = MLP(\hat{O}^l) + O^l \tag{2}$$

where *MSA* represents the multi-headed self-attention mechanism module O^l and O^{l-1} represent the output of the transformer module, *Liner_nor* and *MLP* denote normalization and Multi-Layer Perceptron, we determine the self-attention using:

$$A = Softmax(Q \times K/d^{\frac{1}{2}}) \tag{3}$$

$$Attention = A \times V \tag{4}$$

where Q, K, V denote queries, keys and values respectively, d is the size of the query and key.

3) JUMP FEATURE FUSION MODULE

As illustrated in Figure 2, the top layer comprises high-resolution features F_{high}, while the remaining layers consist of low-resolution features x_i. The resolution decreases with smaller values of i. Since high-level features are derived from lower levels, adjacent features exhibit similarity. To perform feature fusion, we employ features with diverse resolutions. To ensure feature size consistency, we initially measure all hierarchical features for upsample. Subsequently, for expanding the receptive field, we utilize a 1 × 1 expansion convolution to generate a new feature map, as depicted by the specific formula below:

$$X_i = Conv_{1\times1,dilation=i}(x_i) \tag{5}$$

where i is the number of layers of the feature, After that, concatenate the output results of layer $i + 3$ and pass 1 × 1 convolutional block for channel count reduction.

$$Y_i = Conv_{1\times1}[Concat\ (X_i, X_{i+3})] \tag{6}$$

where *Concat* is concatenate, the new high-resolution feature map Z is then obtained by Concatenation after Squeeze processing of all processed features:

$$Z = Concat[Squeeze\ (X_i), \ldots, Squeeze\ (X_n)] \tag{7}$$

where *Squeeze* represents using 1 × 1 convolution to restore the channel. Finally, the feature map Z is fused with the high-resolution feature map F_{high},and the output shape is restored through a 1 × 1 convolutional block.:

$$F = Conv_{1\times1}[Concat\ (Z, F_{high})] \tag{8}$$

D. LOSS FUNCTION

We use a combined loss function for pixel-level segmentation, region-block segmentation to constrain the model optimization direction and further improve the segmentation results. The loss is given by:

$$Loss_{Dice} = 1 - \frac{2\,|P \cap T|}{|p| + |T|} \tag{9}$$

$$Loss_{BCE} = -\left[T \log(P) + (1 - T) \log(1 - P)\right] \tag{10}$$

$$Loss = Loss_{Dice} + wLoss_{bce} \tag{11}$$

where T represents the ground truth, P represents the segmentation result, w represents the weight of L_{bce}, and w is taken as 0.5 in this work.

E. IMPLEMENTATION DETAILS

The proposed TransU^2-Net ran on the sever with the suggested framework to function: one 12-core Intel 12700K CPU, one NVIDIA 3080Ti (12GB) GPU, 64GB RAM, CUDA 11.7 + Torch v1.10.2. All experimental and comparison models do not use any pre-trained models already trained. The models are trained using the Adam optimizer with an initial learning rate of 3×10^{-4}, the batch size to 4, The weight decay rate is 10^{-5}, and the momentum is set 0.9. and the training epoch is set to 200.

III. EXPERIMENTS AND RESULT

A. EVALUATION METRICS

We used five standard metrics for measuring the effectiveness of medical segmentation to assess the model's performance in the two datasets. The most used metric in medical image contests is Dice coefficient metric. It is applied to determine how similar two samples are, and crucial to obtain fine-grained information from the border for medical picture segmentation. Hausdorff distance (HD) is sensitive to the segmented border; Dice coefficient metric is sensitive to the interior filling of the mask. In order to evaluate the model segmentation performance, we also use the positive predictive value (PPV) and Sensitivity, for the auxiliary evaluation. The following is the calculating formula:

$$Dice = \frac{2|p \cap t|}{|p| + |t|} \tag{12}$$

$$HD\ (p, t) = max\ \{h\ (p, t)\ , h\ (t, p)\} \tag{13}$$

$$h\ (p, t) \underset{a\in p}{max} = \left\{ \underset{b\in t}{min} \parallel a - b \parallel \right\} \tag{14}$$

$$h\ (p, t) \underset{a\in t}{max} = \left\{ \underset{b\in p}{min} \parallel b - a \parallel \right\} \tag{15}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{16}$$

$$PPV = \frac{TP}{TP + FP} \tag{17}$$

TABLE 3. Comparison results of the proposed method on the BraTS2021 dataset.

Methods	Dice Score (%)↑			Hausdorff Dist. (mm)↓			PPV Score (%)↑			Sensitivity Score (%)↑		
	WT	TC	ET	WT	TC	ET	WT	TC	ET	WT	TC	ET
FCN8s	84.08	79.63	74.27	2.9317	1.9308	0.7601	88.67	89.08	76.01	85.24	83.89	72.25
U-Net	89.09	82.77	77.97	2.7291	1.8205	0.8256	92.16	89.71	82.56	90.19	81.44	76.46
DenseU-Net	91.20	84.94	84.80	0.3716	0.7844	0.8927	94.39	92.60	**89.27**	91.46	83.21	83.27
U-Net++	91.30	84.61	84.21	0.2517	0.6100	0.8751	95.09	91.11	87.51	89.06	83.81	83.93
U-Net3+	91.17	83.67	82.21	**0.2007**	0.6776	0.8586	94.39	90.06	85.86	89.30	83.33	82.15
3DU-Net	92.06	85.61	84.50	0.3047	0.7418	0.8533	92.40	88.76	85.53	92.34	87.52	**87.05**
3DV-Net	92.05	85.60	84.65	0.3296	0.7358	0.8462	91.28	87.81	84.62	**93.85**	**88.63**	88.41
U^2-Net	91.82	84.96	84.08	0.2567	0.6517	0.6794	**95.06**	90.79	87.43	89.85	84.74	83.74
CA-Net	90.82	\	\	\	\	\	\	\	\	\	\	\
Our	**92.30**	**86.32**	**85.88**	0.2976	**0.5370**	**0.4734**	94.79	**91.80**	88.62	90.83	86.43	85.77

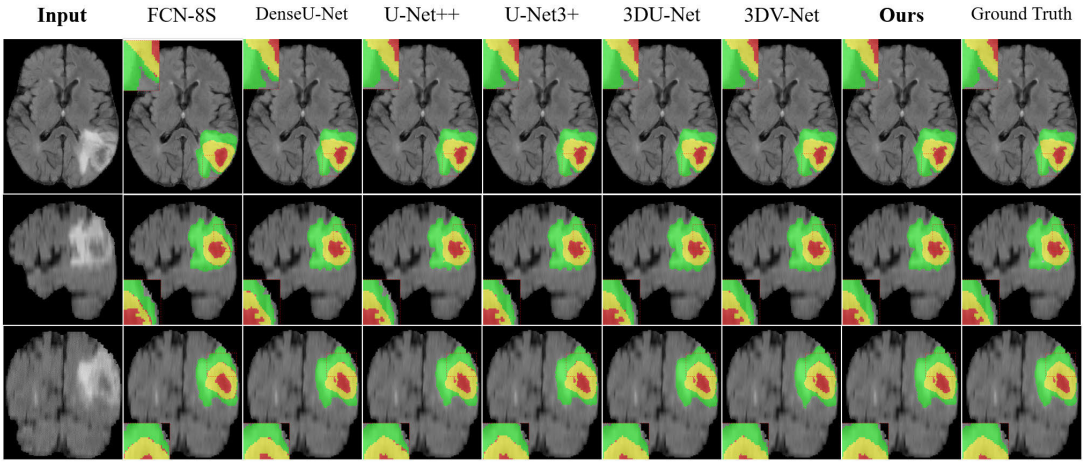

FIGURE 5. Segmentation results on the BraTS2021 dataset.

$$Jaccard = \frac{|p \cap t|}{|p \cup t|} \tag{18}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{19}$$

where p represents the ground truth, t represents the segmentation result. The $\| \cdot \|$ is the distance normal form between point sets ground truth and segmentation result; TP, TN, FP and FN indicate true-positive, true negative, false-positive, and false-negative predictions.

B. MAIN RESULTS

1) BraTS2021 DATASET

To demonstrate the overall segmentation performance of TransU^2-Net, we compare it with other excellent segmentation directions. We have evaluated TransU^2-Net with four types of methods, covering one 2D convolutional segmentation-based method: FCN [6], five U-Net based methods: U-Net [7], DenseU-Net [35], U-Net++ [10], U-Net3+ [36], U^2-Net [37]; two 3D convolution-based segmentation methods: 3DU-Net [13], 3DV-Net [12]; and one kind of transformer-based segmentation methods CA-Net [38]. Although the PPV and Sensitivity of TransU^2-Net were not superior to all other methods, the qualitative results still showed their competitiveness. The Dice coefficient of our TransU^2-Net for WT, ET, and TC are 92.30%, 86.32%, and 85.88%. Table 3 shows the quantitative findings, which are equivalent to or better than the 2D/3D approaches indicated in the table. In terms of Hausdorff distance, the disparity between TransU^2-Net and U-Net3+ [36] is marginal, but TransU^2-Net outperforms U-Net3+ in terms of PPV and Sensitivity.

As shown in Figure 5, we visualize the segmentation results of several representative segmentation networks. In Figure 5, it can be seen that, FCN [6], DenseU-Net [35], U-Net++ [10] and U-Net3+ [36] appear to be unclear for boundary segmentation. Compared to the rest of 3DU-Net [13] and 3DV-Net [12], TransU^2-Net has the sharpest borders and less noise. The attention-based method effectively suppresses the irrelevant background region while reducing noise interference to the segmented region. Furthermore, the proposed jump feature fusion module assists in the learning of more complex semantic features for distinguishing objects in complex situations. In Figure 6, we show the 3D segmentation result displayed in BraTS2021 data and can see that the boundary

IEEE Journal of Translational Engineering in Health and Medicine

TABLE 4. Comparison results of the proposed method on the Msd Task01 dataset.

Methods	Dice Score (%)↑		Hausdorff Dist. (mm)↓		PPV Score (%)↑		Sensitivity Score (%)↑	
	WT	TC	WT	TC	WT	TC	WT	TC
FCN8s	68.91	63.59	6.6900	6.7504	69.15	63.56	70.01	64.91
AttU-Net	72.76	67.51	5.7452	5.8766	72.99	67.86	73.86	69.06
U-Net	72.70	67.14	5.7457	5.8504	73.14	67.09	73.68	68.59
DenseU-Net	73.91	67.14	**5.0507**	6.0012	74.84	67.27	73.86	69.04
U-Net++	72.93	66.96	5.1663	**5.2713**	71.64	64.35	75.59	71.20
U-Net3+	74.69	69.72	5.0890	6.8498	75.11	71.38	75.78	**71.32**
Our	**74.80**	**70.06**	5.1921	7.0084	**75.47**	**71.51**	**75.97**	65.98

TABLE 5. Comparison results of the proposed method on the Msd Task03 dataset.

Methods	Dice Score (%)↑		Jaccrad Score (%)↑		PPV Score (%)↑		Sensitivity Score (%)↑		Accuracy Score (%)↑	
	Liver	Tumor	Liver	Tumor	Liver	Tumor	Liver	Tumor	Liver	Tumor
U-Net	96.18	35.18	92.70	25.90	96.07	46.34	**96.34**	45.21	99.58	99.87
DenseU-Net	95.91	36.70	92.23	27.89	96.17	48.48	95.74	46.27	99.55	99.87
AttU-Net	**96.29**	38.62	**92.91**	29.12	96.43	48.50	96.20	**50.08**	99.59	**99.88**
U-Net++	96.25	36.72	92.83	27.77	96.43	45.41	96.11	49.00	99.58	99.87
U-Net3+	96.17	35.88	92.69	26.95	96.46	48.55	95.92	44.90	99.58	99.87
TransU-Net	95.92	35.31	92.22	26.41	95.92	43.06	95.98	47.22	95.55	99.87
Our	96.24	**42.13**	92.81	**32.69**	**96.47**	**61.65**	96.04	47.45	**99.59**	99.87

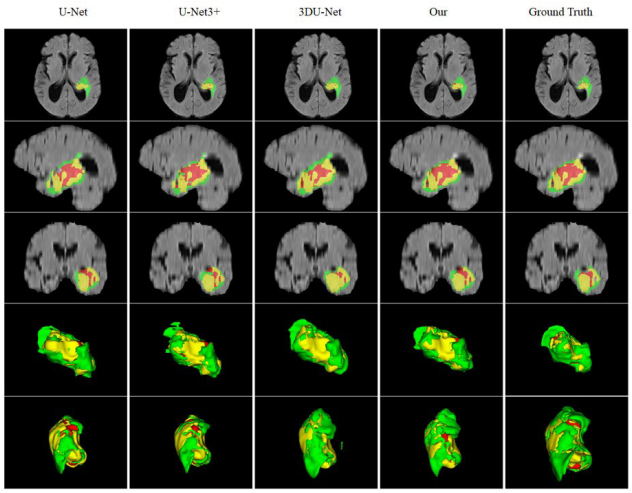

FIGURE 6. Display of 3D segmentation results. WT (green), TC (yellow), ET (red).

of the model segmented by our method is more precise and less noisy. As a result, our model can describe more details, resulting in the best visual segmentation performance.

2) MEDICAL SEGMENTATION DECATHLON DATASET

We evaluated most of the mainstream 2D segmentation models in MSD Task01 data, including FCN [6], U-Net [7], AttU-Net [39], DenseU-Net [35], U-Net++ [10], U-Net3+ [36], six segmentation models. TransU^2-Net achieved a better segmentation, and tumor core are 74.69% and 69.72%, respectively in the non-large medical image dataset of MSD; in the PPV metric, the WT and TC metrics improved by 0.36% and 0.13%, respectively, relative to the best-performing U-Net3+ [36], and the WT also reached the best in the sensitivity metric. Which also has a significant improvement for Hausdorff distance. The numerical outcomes are displayed in Table 4.

As shown in Figure 7, we show the visualization of several models in the MSD dataset, and we can see that our segmentation results are closer to ground truth than U-Net [7] and AttU-Net [39], and TransU^2-Net combines the advantages of transformer and convolution, which can segment brain tumors more accurately and get closer to the factual results.

To validate the effectiveness of TransU^2-Net in other medical image segmentation tasks, we applied the model to the MSD Task03_liver segmentation task. A detailed comparison of different models is presented in Table 5. As shown in Table 5, TransU^2-Net proposed in this study achieved superior performance compared to other compared networks, with the highest Dice coefficient attained in the liver tumors segmentation. Compared to TransU-Net [40], which also utilizes a transformer, TransU^2-Net, featuring a deep U-Net structure as an encoder, is capable of extracting deeper-level features more effectively. The use of the transformer facilitates the retention of location information while capturing global contextual information, surpassing other attention mechanism-based networks.

In Figure 8, we present the segmentation results for the MSD task03 liver using TransU^2-Net. Our results indicate that TransU^2-Net performs exceptionally well with small datasets and outperforms other networks by producing fewer false segmentations. This superior performance can be attributed to the unique combination of the deep U-Net's

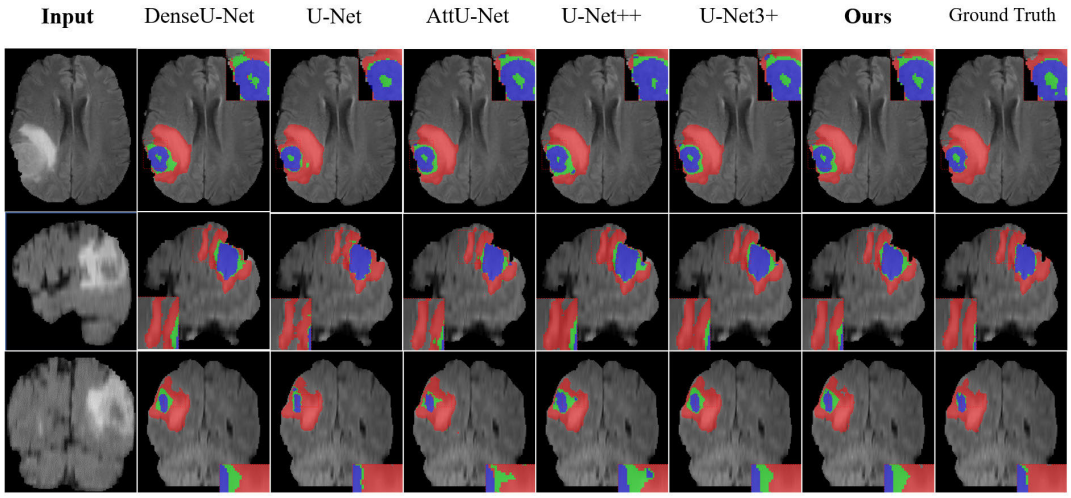

FIGURE 7. Segmentation results on the MSD Task01 dataset.

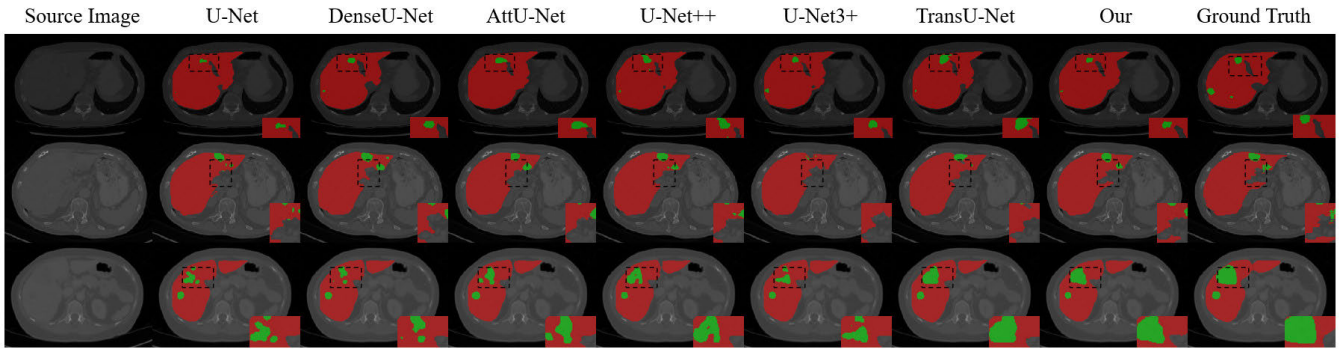

FIGURE 8. Segmentation results on the MSD Task03 dataset.

TABLE 6. Impact of the transformer module on different segmentation networks in the Msd Task03 dataset.

Methods	AVG Dice Score (%)↑	AVG Jaccrad Score. (mm)↑	AVG PPV Score (%)↑	AVG Sensitivity Score (%)↑
U-Net	65.68±43.13	59.30±47.23	71.21±35.16	70.78±36.15
DenseU-Net	66.31±41.86	60.06±45.49	72.33±33.72	70.01±36.01
AttU-Net	67.46±40.78	61.02±45.11	72.47±33.89	73.14±32.61
U^2-Net	66.21±42.21	60.03±45.87	72.69±33.01	68.79±34.41
Add Transformer to U-Net	65.96±42.61	59.50±46.70	73.86±31.19	69.95±37.26
Add Transformer to DenseU-Net	67.87±40.04	61.78±43.69	74.66±30.68	71.19±35.16
Add Transformer to AttU-Net	67.73±40.31	62.24±43.21	72.95±32.85	73.45±32.33
Add Transformer to U^2-Net	68.17±39.51	61.82±43.54	78.98±24.87	69.81±36.65

feature extraction capability and the transformer's ability to capture global features.

C. ABLATION STUDY

1) IMPACT OF THE TRANSFORMER MODULE

To validate the effectiveness of our proposed approach that combines deep convolution with transformer, we incorporate the transformer module into various segmentation networks and evaluate their performance. The experimental results in Table 6 demonstrate that adding the transformer module to all segmentation networks leads to significant improvements, with the U^2-Net achieving the highest boost in AVG Dice coefficient (1.96). These experiments confirm that the combination of a deep U-shaped network with a transformer module is a powerful approach for medical image segmentation.

To further illustrate the advantages of TransU^2-Net, we conducted an evaluation of the significance of the converter module in the segmentation task. Our experimental results, as presented in Table 7, indicate that the transformer

IEEE Journal of Translational Engineering in Health and Medicine

TABLE 7. Impact of the transformer module on the BraTS2021 dataset.

Methods	AVG Dice Score (%)↑	AVG Hausdorff Dist. (mm)↓	AVG PPV Score (%)↑	AVG Sensitivity Score (%)↑
No Transformer model	87.08±4.16	0.5821±0.28	90.98±4.04	86.39±3.09
Add Transformer to the encoder	87.10±3.82	0.5115±0.13	91.40±3.12	86.22±3.60
Add Transformer to skip connect	87.90±3.89	0.4121±0.10	92.70±2.26	86.61±3.84

TABLE 8. Impact of the jump feature fusion module on the BraTS2021 dataset.

Methods	AVG Dice Score (%)↑	AVG Hausdorff Dist. (mm)↓	AVG PPV Score (%)↑	AVG Sensitivity Score (%)↑
Output last feature map	86.24±5.19	0.6129±0.25	91.16±1.09	85.47±6.86
Concat feature maps	86.95±4.24	0.5229±0.23	91.01±4.24	85.77±4.02
Add jump feature fusion	87.42±4.08	0.5048±0.20	91.32±3.91	86.46±3.34

module plays a crucial role in feature extraction. Specifically, the transformer module contributed to a considerable increase in PPV (0.42%), and a decrease in Hausdorff distance (0.07). In addition, by incorporating the transformer module into the jump connection and increasing network depth, we observed improved fusion of global and local features and better handling of information loss, as reflected by a significant increase in Dice coefficient (0.80%) and a reduction in Hausdorff distance (0.10). These findings suggest that a reasonable combination of deeply nested U-shaped structures and transformers can be highly effective in segmenting target lesions.

2) ABLATION ON JUMP FEATURE FUSION MODULE

Finally, we assess the jump feature fusion module overall segmentation performance on the model. The test results in Table 7 demonstrate that our jump feature fusion module is crucial in feature fusion. jump feature fusion module explicitly gives the model an improvement of 1.18 Dice coefficient. We combine features from various levels using the jump feature fusion module, resulting in segmented visuals with the spatial information of high-level features and the semantic information of low-level features. The outcomes demonstrate that segmenting targets using the jump feature fusion module is advantageous.

IV. CONCLUSION

In this paper, we propose a novel automatic segmentation method for multi-modality brain tumor segmentation in MRI based on a deeply nested U-shaped structure by combining transformer with jump feature map fusion. The final architecture not only inherits the advantages of transformer in learning global semantic associations but also uses different levels of features to make the model retain more semantics and more details. TransU^2-Net has advantages in learning global semantic associations and employing different levels of features, allowing the model to retain more semantics and details. The results of this paper proposed model TransU^2-Net on three datasets validate its effectiveness.

In the future, we will improve and expand the TransU^2-Net architecture to extend the model to 3D segmentation. Will also consider introducing the domain adaptation method, so that the model can be adapted to different modalities and achieve optimal segmentation in terms of device parameters.

DATA AVAILABILITY STATEMENT

The BraTS2021 data that support this study are openly available at https://www.med.upenn.edu/cbica/brats2021; MSD data that support this study are openly available at http://medicaldecathlon.com.

REFERENCES

[1] H. Xu et al., "ITGB2 as a prognostic indicator and a predictive marker for immunotherapy in gliomas," *Cancer Immunol., Immunotherapy*, vol. 71, no. 3, pp. 645–660, Mar. 2022.

[2] A. Crimi, S. Bakas, H. Kuijf, B. Menze, and M. Reyes, Eds., *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries Third International Workshop, BrainLes 2017, Held in Conjunction With MICCAI 2017, Quebec City, QC, Canada, September 14, 2017, Revised Selected Papers*, vol. 10670. Cham, Switzerland: Springer, Sep. 2017.

[3] T. Zhou, S. Ruan, and S. Canu, "A review: Deep learning for medical image segmentation using multi-modality fusion," *Array*, vols. 3–4, Sep. 2019, Art. no. 100004.

[4] J. Li et al., "TransBTSV2: Towards better and more efficient volumetric segmentation of medical images," 2022, *arXiv:2201.12785*.

[5] H.-J. Yoo, "Deep convolution neural networks in computer vision: A review," *IEIE Trans. Smart Process. Comput.*, vol. 4, no. 1, pp. 35–43, Feb. 2015.

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

[8] A. M. Nazif and M. D. Levine, "Low level image segmentation: An expert system," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-6, no. 5, pp. 555–577, Sep. 1984.

[9] J. Wang et al., "Coarse-to-fine multiplanar D-SEA UNet for automatic 3D carotid segmentation in CTA images," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 16, no. 10, pp. 1727–1736, Oct. 2021.

[10] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, Jun. 2020.

[11] C. Kou, W. Li, Z. Yu, and L. Yuan, "An enhanced residual U-Net for microaneurysms and exudates segmentation in fundus images," *IEEE Access*, vol. 8, pp. 185514–185525, 2020.

[12] 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.

[13] 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. Intervent.* Cham, Switzerland: Springer, Oct. 2016, pp. 424–432.

[14] M. Havaei et al., "Brain tumor segmentation with deep neural networks," *Med. Image Anal.*, vol. 35, pp. 18–31, Jan. 2017.

[15] F. Lu, F. Wu, P. Hu, Z. Peng, and D. Kong, "Automatic 3D liver location and segmentation via convolutional neural network and graph cut," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 12, no. 2, pp. 171–182, Feb. 2017.

[16] P. Moeskops, M. A. Viergever, A. M. Mendrik, L. S. de Vries, M. J. N. L. Benders, and I. Išgum, "Automatic segmentation of MR brain images with a convolutional neural network," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1252–1261, May 2016.

[17] S. Valverde et al., "Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach," *NeuroImage*, vol. 155, pp. 159–168, Jul. 2017.

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

[19] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–15.

[20] Y. Zhang, H. Liu, and Q. Hu, "Transfuse: Fusing transformers and CNNs for medical image segmentation,"in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, Sep. 2021, pp. 14–24.

[21] W. Wang et al., "TransBTS: Multimodal brain tumor segmentation using transformer," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2021, pp. 109–119.

[22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[23] Z. Wu, C. Shen, and A. Van Den Hengel, "Wider or deeper: Revisiting the ResNet model for visual recognition," *Pattern Recognit.*, vol. 90, pp. 119–133, Jun. 2019.

[24] E. Xie et al., "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inf. Process. Sys. (NIPS)*, vol. 34, Dec. 2021, pp. 12077–12090.

[25] S. Zheng et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6877–6886.

[26] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7242–7252.

[27] Y. Xie, J. Zhang, C. Shen, and Y. Xia, "CoTr: Efficiently bridging CNN and transformer for 3D medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2021, pp. 171–180.

[28] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.

[29] H. Wu, J. Zhang, K. Huang, K. Liang, and Y. Yu, "FastFCN: Rethinking dilated convolution in the backbone for semantic segmentation," 2019, *arXiv:1903.11816*.

[30] G. Li and Y. Yu, "Visual saliency detection based on multiscale deep CNN features," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5012–5024, Nov. 2016.

[31] B. H. Menze et al., "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE Trans. Med. Imag.*, vol. 34, no. 10, pp. 1993–2024, Oct. 2015.

[32] S. Bakas et al., "Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features," *Sci. Data*, vol. 4, no. 1, pp. 1–13, Sep. 2017.

[33] S. Bakas et al., "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge," 2018, *arXiv:1811.02629*.

[34] A. L. Simpson et al., "A large annotated medical image dataset for the development and evaluation of segmentation algorithms," 2019, *arXiv:1902.09063*.

[35] A. Kaku et al., "DARTS: DenseUnet-based automatic rapid tool for brain segmentation," 2019, *arXiv:1911.05567*.

[36] 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.

[37] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, "U2-Net: Going deeper with nested U-structure for salient object detection," *Pattern Recognit.*, vol. 106, Oct. 2020, Art. no. 107404.

[38] R. Gu et al., "CA-net: Comprehensive attention convolutional neural networks for explainable medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 40, no. 2, pp. 699–711, Feb. 2021.

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

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