IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 9 October 2023, accepted 2 November 2023, date of publication 8 November 2023, date of current version 13 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3330958

RESEARCH ARTICLE

TransDoubleU-Net: Dual Scale Swin Transformer With Dual Level Decoder for 3D Multimodal Brain Tumor Segmentation

MARJAN VATANPOUR AND JAVAD HADDADNIA
Department of Biomedical Engineering, Hakim Sabzevari University, Sabzevar 96186-76115, Iran
Corresponding author: Javad Haddadnia (haddadnia@hsu.ac.ir)

ABSTRACT Segmenting brain tumors in MR modalities is an important step in treatment planning. Recently, the majority of methods rely on Fully Convolutional Neural Networks (FCNNs) that have acceptable results for this task. Among various networks, the U-shaped architecture known as U-Net, has gained enormous success in medical image segmentation. However, absence of long-range association and the locality of convolutional layers in FCNNs can create issues in tumor segmentation with different tumor sizes. Due to the success of Transformers in natural language processing (NLP) as a result of using self-attention mechanism to model global information, some studies designed different variations of vision based U-Shaped Transformers. So, to get the effectiveness of U-Net we proposed TransDoubleU-Net which consists of double U-shaped nets for 3D MR Modality segmentation of brain images based on dual scale Swin Transformer for the encoder part and dual level decoder based on CNN and Transformers for better localization of features. The model's core uses the shifted windows multi-head self-attention of Swin Transformer and skip connections to CNN based decoder. The outputs are evaluated on BraTS2019 and BraTS2020 datasets and showed promising results in segmentation.

INDEX TERMS Brain tumor segmentation, vision transformer, Swin transformer, dual scale, U-Net, BraTS.

I. INTRODUCTION

Computer vision has been used for analysis of medical images as a result of the advancement of deep learning. Image segmentation is essential in medical image analysis since it is frequently the initial step in analyzing anatomical structures [1]. After the introduction of U-shaped network, U-Net [2], convolutional neural networks (CNNs) were chosen as the main method to do this task [3], [4]. Further, the performance of the U-Net was improved by some variants of it such as Res-UNet [5], [6], 3D U-Net [7] and U-Net++ [8]. Although CNN-based methods play a major role in segmenting medical images, they also have limitations. First, the convolution only collects data from nearby pixels and cannot explicitly express long-range and global semantic information interaction. Second, because convolution kernels are often fixed in size and shape, they cannot adapt to the input content [9].

By combining the CNN methods with the attention mechanism [10], [11], [12], using atrous convolutional layer [13], [14] and image pyramids [15], [16], these restrictions can be overcome to some extent. These methodologies, however, have drawbacks when it comes to representing long-range dependencies.

Recently, Transform-based frameworks have been proposed as an alternate design, and it has performed well on a variety of computer vision tasks. Chen et al. [17] used the strong combination of Transformers and U-Net as an alternative method for segmenting medical images. Since Transformer training requires large-scale datasets, Valanarasu et al. [18] proposed a gated position-sensitive axial attention mechanism and Local-Global methodology for training the Transformer which was successful. Cao et al. [19] concluded that using skip connections for Transformer

The associate editor coordinating the review of this manuscript and approving it for publication was Vishal Srivastava.

is effective, so they constructed an encoder-decoder architecture using skip connections based on Swin Transformer. For extraction of the volumetric spatial feature maps, 3D CNN is used in the encoder by Wang et al. [20] and Transformer is utilized to model global features.

Gao et al. [9] suggested a U-shape hybrid Transformer network, which combines improvements in convolutional layers with a self-attention mechanism. The hybrid approach enables Transformer to be initialized into Conv-Nets without the need for pre-training and self-attention enables operations in encoder and decoder layers effectively achieve global dependencies. Vanilla Transformer treats equally each image position, but to reduce the computational costs and focus only on special part of image, Xie et al. [21] introduced a new attention mechanism in which only some key parts around a reference point are considered for the self-attention mechanism. For segmentation of 3D images, Hatamizadeh et al. [22] proposed a method that learns representations of the input with the help of a Transformer as the encoder.

The Transformer's self-attention mechanism simply investigates the feature maps spatial dimension connections, with no interaction with the channel dimension. So, Wu et al. [23] developed Dimensional Interactive (DI) self-attention that its combination with U-Net (DI-Unet) had better performance than Swin-Unet in large datasets.

Inspired by these observations, in this study we have proposed a novel approach for brain tumor segmentation and our original contributions are:

- ☐ TransDoubleU-Net as a novel architecture combining a dual-scale Swin Transformer.
- ☐ Designing a dual-level decoder which consist of convolutional layers and transformers.
- ☐ Combining the encoder with dual-level decoder using skip connections and Transformers.

II. MATERIAL AND METHODS

A. ARCHITECTURE OVERVIEW

The architecture of the proposed approach is illustrated in Figure 1. The input has a size of H×W×D×4 as we have 4 modalities for every subject. The proposed method consists of double U-Shaped networks which we call "TransDoubleU-Net" that every part is described as below.

1) NETWORK ENCODER

The raw 3D image is fed to a pre-processing unit. After removing the backgrounds and resizing, the output is in size of H×W×D×4. The encoder's main part is a Dual-Swin Transformer (Figure 2). In order to create the appropriate tokens for Swin Transformer, a patch partition layer is considered. The feature dimension of patch partitioning layer is 256 and 32 with patch resolutions of $(\frac{H}{2S})^3$ and $(\frac{H}{S})^3$ respectively, where s denotes the patch size. A linear embedding layer is creating 3D tokens to Swin Transformer with dimensions of C=48 for patches of size S and 2C = 96 for patches of size *2S*. The self-attention mechanism is computed in non-overlapping windows of size M=8. In every subsequent layer of the Transformer the windows are shifted by M/2. We can calculate the outputs of every stage in layer l and $l+1$ with the following:

$$\begin{aligned}
\hat{z}^l &= W - MSA\left(LN\left(z^{l-1}\right)\right) + z^{l-1},\\
z^l &= MLP\left(LN\left(\hat{z}^l\right)\right) + \hat{z}^l,\\
\hat{z}^{l+1} &= SW - MSA\left(LN\left(z^l\right)\right) + z^l,\\
z^{l+1} &= MLP\left(LN\left(z^{l+1}\right)\right) + \hat{z}^{l+1},
\end{aligned} \tag{1}$$

The outputs of *W-MSA* and *SW-MSA* are windowed multi-head self-attention and shifted window denoted as $\hat{z}^l$ and $\hat{z}^{l+1}$. The input for every windowed multi-head self-attention is passed through an *LN* and an *MLP* which they represent layer normalization and multilayer perceptron respectively. The *W-MSA* applies self-attention in windows. To overcome the problem of effective information between windows the *SW-MSA* will help. The input for *SW-MSA* will pass through a *LN* first. To compute the shifted window, a 3D cyclic-shifting [22] is considered and the attention is computed as below:

$$Attention(QKV) = SoftMax(QK^T/\sqrt{d} + B) * V \tag{2}$$

where *Q, K, V* denote the query, key and values. *d* and *B* show dimension of query or key and bias matrix, respectively.

For denoting a hierarchical representation, the tokens are reduced in every stage of the network, but the size of features are doubled with the help of a patch merge layer. Patch merge layer concatenates features of dimension 2 × 2×2 and then applying a linear layer to reduce the feature dimension by factor of 2. So the output resolution is reduced by a factor of 2 and the feature dimension is increased by a factor of 2 for every stage of the Swin Transformer. So, the output of every stage is $(\frac{H}{S})$, $(\frac{H}{2S})$, $(\frac{H}{4S})$ and $(\frac{H}{8S})$ for one Swin Transformer with patch size of C and $(\frac{H}{2S})$, $(\frac{H}{4S})$, $(\frac{H}{8S})$ and $(\frac{H}{16S})$ for Swin Transformer with patch size of 2C.

2) NETWORK DECODER

The decoders' blocks are shown in Figure 3 the decoder connects to encoder with the help of skip connections. For every connection a block named B consists of convolution layers is used for creating richer representation of features. The feature resolution is doubled and the number of features are halved with the help of block M. The 1 × 1 × 1 convolution layer helps with controlling the features dimension. From bottom to top of the decoder feature dimensions are halved except for the top layer where only the resolution of features are doubled for the direction with patch embedding size of C and quadrupled for patch embedding size of 2C.

In the next part, after concatenation of the related parts with skip connections, the first, two bottom layers are concatenated properly through K and L blocks and converted to desired shape with block R, then fed to a vanilla Transformer. The same procedure is applied to the second, two layers

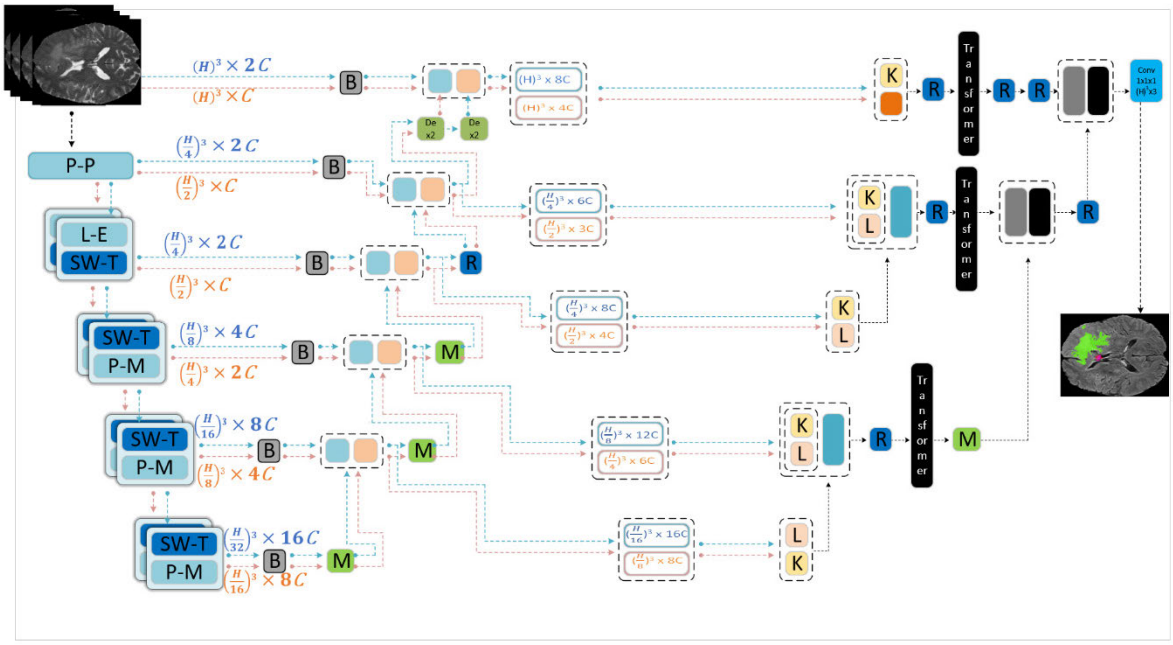

FIGURE 1. Overview of TransDoubleU-Net. P-P, L-E and P-M stand for Patch Partitioning, Linear Embedding and Patch Merging, respectively. The architecture is consist of a double Swin transformer as the encoder. The decoder connects to encoder with the help of skip connections. The output is created by combining the transformer results of different layers.

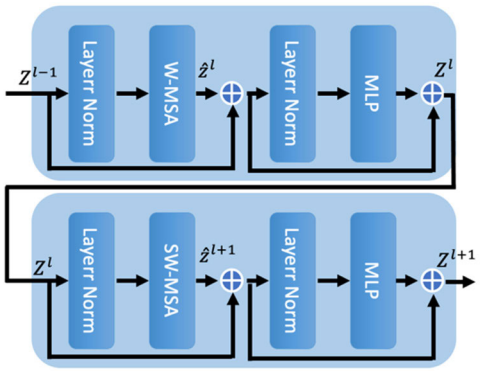

FIGURE 2. The architecture of Swin Transformer block. The W-MSA and SW-MSA are windowed multi-head self-attention and shifted window.

from bottom of the architecture and the results are concatenated with the last step. The top most layer goes through a Transformer and after proper dimension reduction, it is concatenated with the last step result. The final segmented output is calculated via a $1 \times 1 \times 1$ convolution layer.

B. LOSS FUNCTION

The loss function that is used in this study is calculated as the sum of weighted cross-entropy and soft Dice loss function [24]:

$$L_{CE} = -\sum_{i=1}^{N} y_i log\hat{y}_i \tag{3}$$

$$L_{DL}(Y, \hat{Y}) = 1 - \frac{2}{N}\sum_{n=1}^{N}\frac{\sum_{k=1}^{K} Y_{k,n}\hat{Y}_{k,n}}{\sum_{k=1}^{K} Y_{k,n}^2 + \sum_{k=1}^{K}\hat{Y}_{k,n}^2} \tag{4}$$

$$L_{total} = \alpha L_{CE} + \beta L_{DL} \tag{5}$$

where the number of classes is denoted as N, K is voxel number, $Y_{k,n}$ and $\hat{Y}_{k,n}$ represents one-hot encoded label and the probability of output for class n at voxel k, respectively.

III. RESULTS

A. DATASET

The datasets that are used in this work are provided by the Brain Tumor Segmentation: BraTS2019 and BraTS2020 challenge [25], [26], [27], [28]. There are 335 and 369 patients for training in BraTS2019 and BraTS2020, respectively and both datasets consist of 125 cases for validation. Four modalities of brain MRI scans are included in each sample, which are named: T1-weighted (T1), post-contrast T1-weighted (T1ce), T2-weighted (T2) and Fluid Attenuated Inversion Recovery (FLAIR). The size of all the MR images is $240 \times 240 \times 155$. The labels of both datasets are as follows: Label 0: Background, Label 1: Necrotic and non-enhancing tumor (NCR/NET), Label 2: Peritumoral edema (ED) and Label 4: GD-enhancing tumor.

B. IMPLEMENTATION DETAILS

The proposed method is done in python 3.8 with PyTorch. The operating system is Windows 10 and CPU is Intel i7 11700k with 32 Gb of RAM and Nvidia RTX 3090 GPU. The model trained with Adam optimizer with initial learning rate of 0.0003 for 50 epochs, but other criteria such as early stopping, etc. were considered.

To increase data diversity the following data augmentation is applied for all training cases:

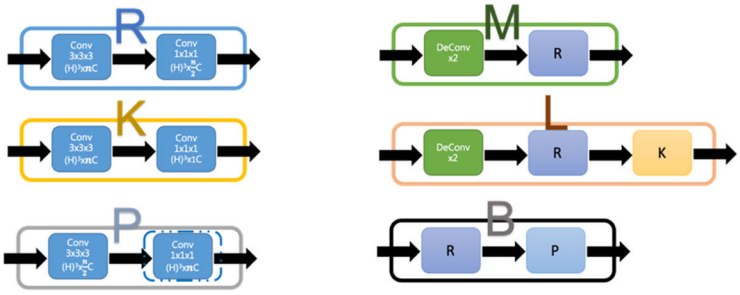

FIGURE 3. The decoders' blocks. The R block is consisted of two Convolutional layers and it is our primary block which is used in other blocks. In layer P, before creating the last output an Instance Normalization is applied. Layers M and L are used to create proper output size with the help of Deconvolutional layers.

TABLE 1. Comparison of SOTA approaches with proposed method on BraTS2019 validation dataset.

Method	DSC %			HD (mm)		
	ET	WT	TC	ET	WT	TC
3D U-Net [29]	70.86	87.38	72.48	5.062	9.432	8.719
V-Net [24]	73.89	88.73	76.56	6.131	6.256	8.705
KiU-Net [36]	73.21	87.60	73.92	6.323	8.942	9.893
Attention U-Net [37]	75.96	88.81	77.20	5.202	7.756	8.258
TransBTS [20]	78.93	90.00	81.94	3.736	5.644	6.049
TransDoubleU-Net	**78.97**	**91.44**	**83.36**	**3.071**	**4.228**	**5.481**

TABLE 2. Comparison of SOTA approaches with proposed method on BraTS2020 validation dataset.

Method	DSC %			HD (mm)		
	ET	WT	TC	ET	WT	TC
3D U-Net [29]	68.76	84.11	79.06	50.983	13.366	13.607
Residual U-Net [30]	71.63	82.46	76.47	37.42	12.34	13.11
Attention U-Net [37]	71.83	85.57	75.96	32.94	11.91	19.43
U-Netr [22]	71.18	88.30	75.85	34.46	8.18	10.63
VTU-Net [31]	76.45	88.73	80.39	28.99	9.54	14.76
TransBTS [20]	78.73	90.09	81.73	17.947	4.964	9.769
SwinBTS [38]	77.36	89.06	80.30	26.84	8.56	15.78
DResU-Net [39]	80.04	86.60	83.57	-	-	-
DenseTrans [32]	**82.30**	91.40	85.30	**15.20**	6.32	16.90
TransDoubleU-Net	79.16	**92.87**	**86.51**	19.953	**4.472**	**7.885**

(1) random flip with probability of 0.5 across all axis, (2) random rotation which randomly rotates the images 0.2 degree, (3) random scaling and (4) random shift intensity was applied on every channel in $[-0.1, 0.1]$ range.

C. EVALUATION METRICS

The segmentation accuracy is evaluated by Dice score (DSC) and Hausdorff distance (95%) metrics. DSC is used to determine the overlap between the ground truth and the predictions and Hausdorff (HD) distance is used to compare ground truth images with segmentation outputs and to rate different segmentation outcomes in medical image segmentation.

$$\mathrm{DSC} = \frac{2TP}{FN + FP + 2TP} \tag{6}$$

$$HD(X, Y) = \max\left\{\sup_{x \in X} \inf_{y \in Y} d(x, y), \sup_{y \in Y} \inf_{x \in X} d(x, y)\right\}, \tag{7}$$

where TP, FN, FP are the number of True Positive, False Negative and False Positive voxels, respectively. And also sup, inf and $d(,)$ denote the supremum, infimum and the function that computes the distance between two points, while x and y represent the points on surface X of the ground truth and surface Y of the predicted regions respectively.

Both metrics are measured for enhancing tumor (ET: Label 1), tumor core (TC: Label 1 + Label 4), and whole tumor (WT: Label 1+ Label 2+ Label 4).

D. QUANTITATIVE RESULTS

We start by doing a five-fold cross-validation evaluation on the BraTS2019 training set, which is a typical configuration used by many studies. Average DSC for our TransDoubleU-Net for ET, WT, and TC are 80.04%, 92.71%, and 85.69%, respectively. Additionally, we do tests on BraTS2019 validation dataset and compared the results

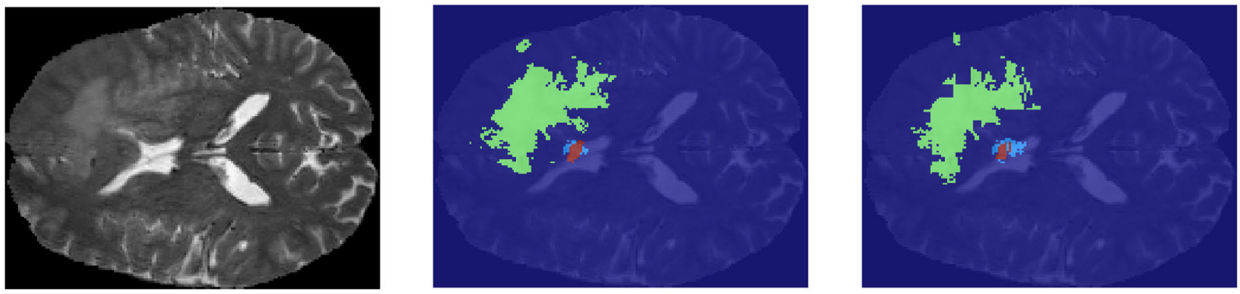

FIGURE 4. Qualitative results of TransDoubleU-Net. WT=Green+Blue+Red, TC=Blue+Red and ET=Blue.

TABLE 3. Ablation study on TransDoubleU-Net architecture.

Model	DSC %		
	ET	WT	TC
Single encoder, Dual decoder	76.64	79.57	80.41
Dual encoder, Single decoder	80.93	85.21	82.30
Dual encoder, Dual decoder	**82.14**	**89.66**	**85.01**

with other existing SOTA approaches in Table 1. The TransDoubleU-Net achieves the DSC of 78.97%, 91.44% and 83.36% for ET, WT and TC, respectively, which outperforms the previous SOTA methods. A significant improvement in segmentation has also been achieved in terms of HD.

We also trained the TransDoubleU-Net on BraTS2020 training dataset and evaluated on validation dataset, the hyperparameters were the same. The comparison results with SOTA models can be seen in Table 2. The TransDoubleU-Net achieves the DSC of 79.16%, 92.87%, 86.51% and HD of 19.953mm, 4.472mm, 7.885mm on ET, WT and TC, respectively.

The quantitative results show that the TransDoubleU-Net has considerable improvement in brain tumor segmentation compared with traditional CNN-based methods such as 3D U-Net [29], V-Net [24] and residual U-Net [30]. Additionally, it outperforms the models that used Transformer structure such as U-Netr [22], VTU-Net [31] and TransBTS [20]. This reveals the benefit of using a Swin Transformer to represent global features and combining with a decoder based on convolution and base Transformer to detect where exactly are the target locations locally.

E. QUALITATIVE RESULTS

The result of TransDoubleU-Net segmentation and ground truth image is shown in Figure 4. The proposed method can achieve prominent results compared to ground truth mask. Although we can see clear errors in some locations, but the difference is acceptable. The produced segmentation mask by TransDoubleU-Net confirms the efficacy of the dual scale Swin Transformer as the encoder and dual decoder for better localization of tumor.

IV. DISCUSSION

The performance of TransDoubleU-Net has been evaluated considering different folds of data against SOTA models. The average DSC and HD for BraTS2019 and BraTS2020 are 84.59%, 4.26mm and 86.18%, 10.77mm, respectively. The results available in Table 1 and Table 2 show that for all classes of ET, WT and TC the TransDoubleU-Net performs very well except ET compared to SOTA [32] in BraTS2020.

Studying winners of BraTS2019 [33] and BraTS2020 [34], the proposed method performs comparable, except in ET in both DSC and HD. It is notable that in BraTS2019 the difference in TC is very small. According to results available in Table 3 and Figure 5, the efficacy of proposed architecture is clear in visual comparison of TransDoubleU-Net and its variation ablation studies. Performance boost of dual scale Swin Transformer is considered meaningful. The dual scale patches fed to Swin Transformers can help for better capturing of long-range dependencies and the windowed multi-head self-attention can embed contextual information of the images. It also helps in reducing the computational complexity of the model by considering windows of size M×M, but in order to capture the information of other windows, it will shift the windows by M/2 in every direction. The different patch sizes of encoder can implicitly induce the combination of local and global view of the inputs. The hierarchy of the encoder part is realized in reducing the resolution in every stage and increasing the feature dimension. The decoder part of the model holds the information of its own stage and the peer encoder stage concatenated together with convolution and instance normalization operations. This approach helps to decode the location of the tumor classes. It is important to notice the second level decoder for more detailed localization

TABLE 4. Ablation study on different C, patch size and levels of transformer.

L	Patch size	C	DSC (%)		
			ET	WT	TC
8	(4 , 8)	24	74.15	83.42	80.48
		48	76.29	85.64	81.93
8	(2 , 4)	24	74.98	84.19	81.17
		48	77.91	86.51	82.32
12	(4 , 8)	24	79.26	87.07	82.86
		48	81.94	88.75	84.22
12	(2 , 4)	24	81.03	87.90	83.54
		48	**82.14**	**89.66**	**85.01**

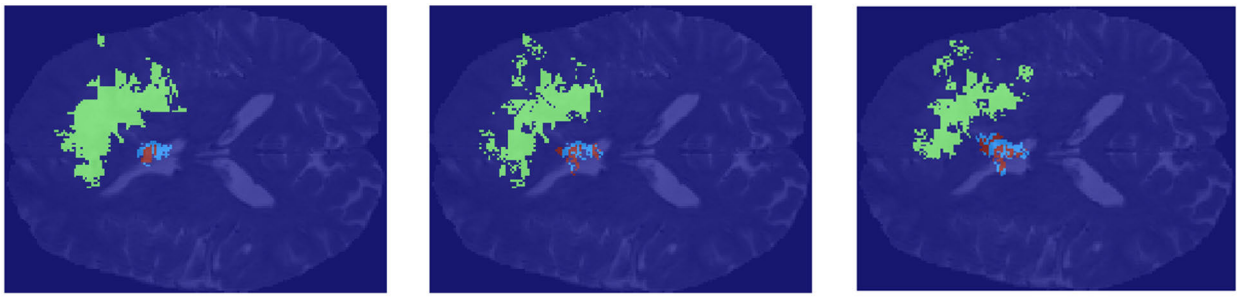

FIGURE 5. Ablation study on TransDoubleU-Net architecture.

TABLE 5. Ablation study on the output results of different decoder layer.

Levels of decoder	DSC %		
	ET	WT	TC
i=0	77.91	86.28	81.76
i=0 + i=(1+2)	80.65	88.63	84.59
i=0 + i=(3+4)	78.80	85.75	80.03
i=0 + i=(1+4)	79.04	86.92	82.80
i=0 + i=(2+3)	80.23	87.84	83.44
i=0 + i=(1+2) + i=(3+4)	**82.14**	**89.66**	**85.01**

TABLE 6. Testing time of different architectural designs in TransDoubleU-Net.

Model	Testing Time (sec)
Single encoder, Dual decoder	0.899
Dual encoder, Single decoder	0.783
Dual encoder, Dual decoder	1.201

of tumor and it is somehow an answer to the dual scale encoder.

For better evaluation of the impact of C and Patch size we performed another study in which different levels of Swin Transformer called L combined with variations of C and Patch size are compared. It is obvious that C=48, L=12 and Patch size = (2,4) can outperform other architectures. The reason for L is obvious as we can say the deeper Swin Transformers get the better results are achieved. Also, the feature dimension C can show better results if we collect more features from Swin Transformers (Table 4).

Another study on second layer of decoder showed that if we combine the output of *i=0* layer of the decoder with other decoder levels, the results may vary slightly and we

chose the best combination which was the {*i=0* + *i=(1+2)* + *i=(3+4)*} (Table 5). In Table 6 the inference time for different architectural designs is available. It can be seen that as the design gets more complex, the results are better in terms of performance but there is a time delay in inference mode. It can induced from the results that the dual decoder can increase the testing time but improve performance results.

V. CONCLUSION

In this study a novel approach called TransDoubleU-Net for 3D semantic segmentation of brain tumor of MR images is proposed based on dual Swin Transformer as an encoder and a dual level decoder based on CNN and Transformers in order to better localization of features. In TransDoubleU-Net we leverage the benefits of hierarchical features of Swin Transformer and combine it with a dual-level decoder. The decoder does not evaluate the results of the encoder not only once but twice and it has proved for better performance. The method was evaluated on BraTS datasets and showed promising results. We believe that TransDoubleU-Net could be a fundamental method for further improvements in medical image segmentation. The method provides better performance in comparison with other related studies. The main reason is because of the dual U-shape networks which work in parallel and leveraging the Swin Transformer benefits. One the main limitations of this study is generalization which is mainly caused by the absence of more concrete datasets. It would be a good practice to provide a more sophisticated dataset for brain tumor segmentation, which we think is doable within near future. Another drawback of this study is testing and training time, which can be addressed in the future works, e.g. by using a modified version of attention, and using more powerful GPUs. For future works it would be of great help to visualize the experimental results of other studies to have a better understanding in terms of visual comparison. Also, we suggest using a modified version of attention such as Scatterbrain [35] for computational complexity reduction by combining sparse and low-rank attention, especially in the decoder part which can also help in reducing the testing time. Nevertheless, It may impact the final performance results but considering the reduction in memory resources dependency, it is a convenient approach. By considering such architecture, we may address some of the drawbacks mentioned before.

AUTHORS' CONTRIBUTIONS

All Authors has contributed equally in this study. All authors read and approved the final manuscript.

REFERENCES

[1] M. Monteiro, V. F. J. Newcombe, F. Mathieu, K. Adatia, K. Kamnitsas, E. Ferrante, T. Das, D. Whitehouse, D. Rueckert, D. K. Menon, and B. Glocker, "Multiclass semantic segmentation and quantification of traumatic brain injury lesions on head CT using deep learning: An algorithm development and multicentre validation study," *Lancet Digit. Health*, vol. 2, no. 6, pp. 314–322, Jun. 2020.

[2] 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, 2015, pp. 234–241.

[3] M. Aghalari, A. Aghagolzadeh, and M. Ezoji, "Brain tumor image segmentation via asymmetric/symmetric UNet based on two-pathway-residual blocks," *Biomed. Signal Process. Control*, vol. 69, Aug. 2021, Art. no. 102841.

[4] G. Du, X. Cao, J. Liang, X. Chen, and Y. Zhan, "Medical image segmentation based on U-Net: A review," *J. Imag. Sci. Technol.*, vol. 64, pp. 1–12, Jan. 2020.

[5] M. Z. Alom, C. Yakopcic, M. Hasan, T. M. Taha, and V. K. Asari, "Recurrent residual U-Net for medical image segmentation," *J. Med. Imag.*, vol. 6, no. 1, 2019, Art. no. 014006.

[6] J. Yang, J. Zhu, H. Wang, and X. Yang, "Dilated MultiResUNet: Dilated multiresidual blocks network based on U-Net for biomedical image segmentation," *Biomed. Signal Process. Control*, vol. 68, Jul. 2021, Art. no. 102643.

[7] W. Chen, B. Liu, S. Peng, J. Sun, and X. Qiao, "S3D-UNet: Separable 3D U-Net for brain tumor segmentation," in *Proc. Int. MICCAI Brainlesion Workshop*. Cham, Switzerland: Springer, 2018, pp. 358–368.

[8] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested U-Net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Berlin, Germany: Springer, 2018, pp. 3–11.

[9] Y. Gao, M. Zhou, and D. N. Metaxas, "UTNet: A hybrid transformer architecture for medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2021, pp. 61–71.

[10] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.

[11] C. Kaul, S. Manandhar, and N. Pears, "FocusNet: An attention-based fully convolutional network for medical image segmentation," in *Proc. IEEE 16th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2019, pp. 455–458.

[12] J. Zhang, Z. Jiang, J. Dong, Y. Hou, and B. Liu, "Attention gate ResU-Net for automatic MRI brain tumor segmentation," *IEEE Access*, vol. 8, pp. 58533–58545, 2020.

[13] R. Arora, B. Raman, K. Nayyar, and R. Awasthi, "Automated skin lesion segmentation using attention-based deep convolutional neural network," *Biomed. Signal Process. Control*, vol. 65, Mar. 2021, Art. no. 102358.

[14] W. Weng and X. Zhu, "INet: Convolutional networks for biomedical image segmentation," *IEEE Access*, vol. 9, pp. 16591–16603, 2021.

[15] S. Feng, H. Zhao, F. Shi, X. Cheng, M. Wang, Y. Ma, D. Xiang, W. Zhu, and X. Chen, "CPFNet: Context pyramid fusion network for medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 10, pp. 3008–3018, Oct. 2020.

[16] H. Park and J. Paik, "Pyramid attention upsampling module for object detection," *IEEE Access*, vol. 10, pp. 38742–38749, 2022.

[17] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.

[18] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, "Medical transformer: Gated axial-attention for medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2021, pp. 36–46.

[19] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-UNet: UNet-like pure transformer for medical image segmentation," 2021, *arXiv:2105.05537*.

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

[21] 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.

[22] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, "UNETR: Transformers for 3D medical image segmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 1748–1758.

[23] Y. Wu, G. Wang, Z. Wang, H. Wang, and Y. Li, "DI-UNet: Dimensional interaction self-attention for medical image segmentation," *Biomed. Signal Process. Control*, vol. 78, Sep. 2022, Art. no. 103896.

[24] F. Milletari, N. Navab, and S.-A. 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.

[25] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani, and C. Davatzikos, "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.
[26] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, R. T Shinohara, C. Berger, S. M. Ha, M. Rozycki, and M. Prastawa, "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge," 2018, *arXiv:1811.02629*.
[27] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, and L. Lanczi, "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE Trans. Med. Imag.*, vol. 34, no. 10, pp. 1993–2024, Oct. 2015.
[28] S. Bakas, A. Sotiras, M. Bilello, M. Rozycki, J. Kirby, J. Freymann, K. Farahani, and C. Davatzikos, "Segmentation labels and radiomic features for the pre-operative scans of the TCGA-LGG collection," Tech. Rep., 2017.
[29] O. 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, 2016, pp. 424–432.
[30] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual U-Net," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 749–753, May 2018.
[31] H. Peiris, M. Hayat, Z. Chen, G. Egan, and M. Harandi, "A robust volumetric transformer for accurate 3D tumor segmentation," 2021, *arXiv:2111.13300*.
[32] L. ZongRen, W. Silamu, W. Yuzhen, and W. Zhe, "DenseTrans: Multimodal brain tumor segmentation using Swin transformer," *IEEE Access*, vol. 11, pp. 42895–42908, 2023.
[33] Z. Jiang, C. Ding, M. Liu, and D. Tao, "Two-stage cascaded U-Net: 1st place solution to brats challenge 2019 segmentation task," in *Proc. Int. MICCAI Brainlesion Workshop*. Shenzhen, China: Springer, 2020, pp. 231–241.
[34] F. Isensee, P. F. Jäger, P. M. Full, P. Vollmuth, and K. H. Maier-Hein, "nnU-Net for brain tumor segmentation," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Lima, Peru: Springer, 2021, pp. 118–132.
[35] B. Chen, T. Dao, E. Winsor, Z. Song, A. Rudra, and C. Ré, "Scatterbrain: Unifying sparse and low-rank attention," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 17413–17426.
[36] J. M. J. Valanarasu, V. A. Sindagi, I. Hacihaliloglu, and V. M. Patel, "KiU-Net: Overcomplete convolutional architectures for biomedical image and volumetric segmentation," *IEEE Trans. Med. Imag.*, vol. 41, no. 4, pp. 965–976, Apr. 2022.
[37] O. Oktay, J. Schlemper, L. Le Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention U-Net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999*.
[38] 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.
[39] 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.

MARJAN VATANPOUR received the M.Sc. degree in biomedical engineering from the Azad University of Mashhad, Mashhad, Iran, in 2015. She is currently pursuing the Ph.D. degree with the Department of Biomedical Engineering, Hakim Sabzevari University, Sabzevar, Iran. Her current research interests include medical image processing, pattern recognition, artificial intelligence, and deep learning.

JAVAD HADDADNIA received the B.Sc. degree in electrical and electronic engineering and the M.Sc. and Ph.D. degrees in electrical engineering from the Amirkabir University of Technology, Tehran, Iran, in 1993, 1996, and 2002, respectively. He has been with the Department of Electrical and Computer Engineering, Hakim Sabzevari University, Sabzevar, Iran, since 2003. He is currently a University Professor. He received the Full University Professorship, in 2017, the Hakim Sabzevari University Award for Excellence in Scholarship Research, and the Creative Activity, from 2003 to 2005, for three years and from 2009 to 2015, for five years. He received the Best Researcher Award from the Ministry of Science, Research and Technology, Iran, in 2006. He was a recipient of the Best Designer Award for National SCADA System for Electrical Power Distribution Network from the Ministry of Energy, Iran, in 2015. Also, he has authored five books, including *The New Trends in the Application of Thermography Science for Diagnostic Purpose* (Supreme Century, USA, 2016) and he has authored more than 160 articles in the artificial intelligence and digital signal processing. His current research interests include digital signal processing, artificial intelligence, machine vision, pattern recognition, and medical engineering.