TCRNET: MAKE TRANSFORMER, CNN AND RNN COMPLEMENT EACH OTHER

*Xinxin Shan[1], Tai Ma[1], Anqi Gu[2], Haibin Cai[3] and Ying Wen[1]**

[1]School of Communication and Electronic Engineering, East China Normal University, Shanghai, China.
[2]School of Computer Science and Technology, East China Normal University, Shanghai, China.
[3]Software Engineering Institute, East China Normal University, Shanghai, China.

ABSTRACT

Recently, several Transformer-based methods have been presented to improve image segmentation. However, since Transformer needs regular square images and has difficulty in obtaining local feature information, the performance of image segmentation is seriously affected. In this paper, we propose a novel encoder-decoder network named TCRNet, which makes **T**ransformer, **C**onvolutional neural network (CNN) and **R**ecurrent neural network (RNN) complement each other. In the encoder, we extract and concatenate the feature maps from Transformer and CNN to effectively capture global and local feature information of images. Then in the decoder, we utilize convolutional RNN in the proposed recurrent decoding unit to refine the feature maps from the decoder for finer prediction. Experimental results on three medical datasets demonstrate that TCRNet effectively improves the segmentation precision.

***Index Terms*—** Transformer-based method, feature information, encoder-decoder network, image segmentation

1. INTRODUCTION

Convolutional neural network (CNN) has made great achievements in medical image segmentation, among which U-Net [1] is a typical network. In recent years, U-Net++ [2], U-Net 3+ [3] and multi-inputs U-Net [4] have optimized the skip connections of U-Net. Chartsias et al. [5] considered both anatomical and modality factors to segment magnetic resonance images (MRIs). Chatterjee et al. [6] combined the random walker algorithm with graph cut optimization to segment brain MRIs. CNN-based methods have small receptive field, while Transformer [7] and vision Transformer (ViT) [8] leverage the attention mechanism to capture global information and extract more representative features. Transformer is often embedded into the encoder and decoder. TransUNet [9] replaces the deepest feature map in the encoder of U-Net with the feature extracted by ViT. TrSeg [10] uses Transformer as a decoder module that receives feature maps extracted by CNN.

However, there are two deficiencies in the above CNN-based and Transformer-based methods. (i) Resizing the original image to be square will cause image distortion. The data source for some methods [2] [3] especially Transformer-based methods [9] demands regular datasets containing square images. In fact, most of the collected images are arbitrary. Therefore, resizing images will cause deformation and disturb the segmentation results. (ii) Image feature information extracted from single network is insufficient. Transformer can retain global information [11], and CNN can recover local spatial information [8]. Since the multi-scale feature maps extracted by CNN can be viewed as sequential data, recurrent neural network (RNN) is very effective to mine the temporal and context information [12] [13]. Thus, it is vital to combine various networks to obtain more comprehensive feature information.

In view of the two deficiencies, in this paper, we propose an encoder-decoder network called TCRNet that makes Transformer, CNN and RNN complement each other to explore the application of medical image segmentation. Firstly, we embed Transformer into the encoder to extract global feature information of the input image, and concatenate it to the local feature information extracted by CNN. Then, in the decoder, we use the convolutional RNN (convRNN) in recurrent decoding unit (RDU) to refine the multi-scale feature maps from Transformer and CNN, and then generate the score maps to finally predict the segmentation result. The major contributions of TCRNet are listed as follows. (i) By the innovative fusion of features extracted by Transformer and CNN, we ensure the complementarity of global and local information to weaken the negative effects of image distortion. (ii) We design a novel RDU to refine the feature maps from Transformer and CNN, which further mines the temporal and context information.

2. METHOD

The structure of the proposed TCRNet is drawn in Fig. 1, which consists of a proposed encoder and a novel RDU decoder. In the encoder, we respectively utilize Transformer and

This work was supported in part by 2030 National Key Research and Development Program of China (2018AAA0100500), the Natural Science Foundation of Shanghai (22ZR1421000), the National Nature Science Foundation of China (no.61773166), Projects of International Cooperation of Shanghai Municipal Science and Technology Committee (14DZ2260800), the Fundamental Research Funds for the Central Universities, and the ECNU Academic Innovation Promotion Program for Excellent Doctoral Students (YBNLTS2021-040). *Corresponding author: ywen@cs.ecnu.edu.cn

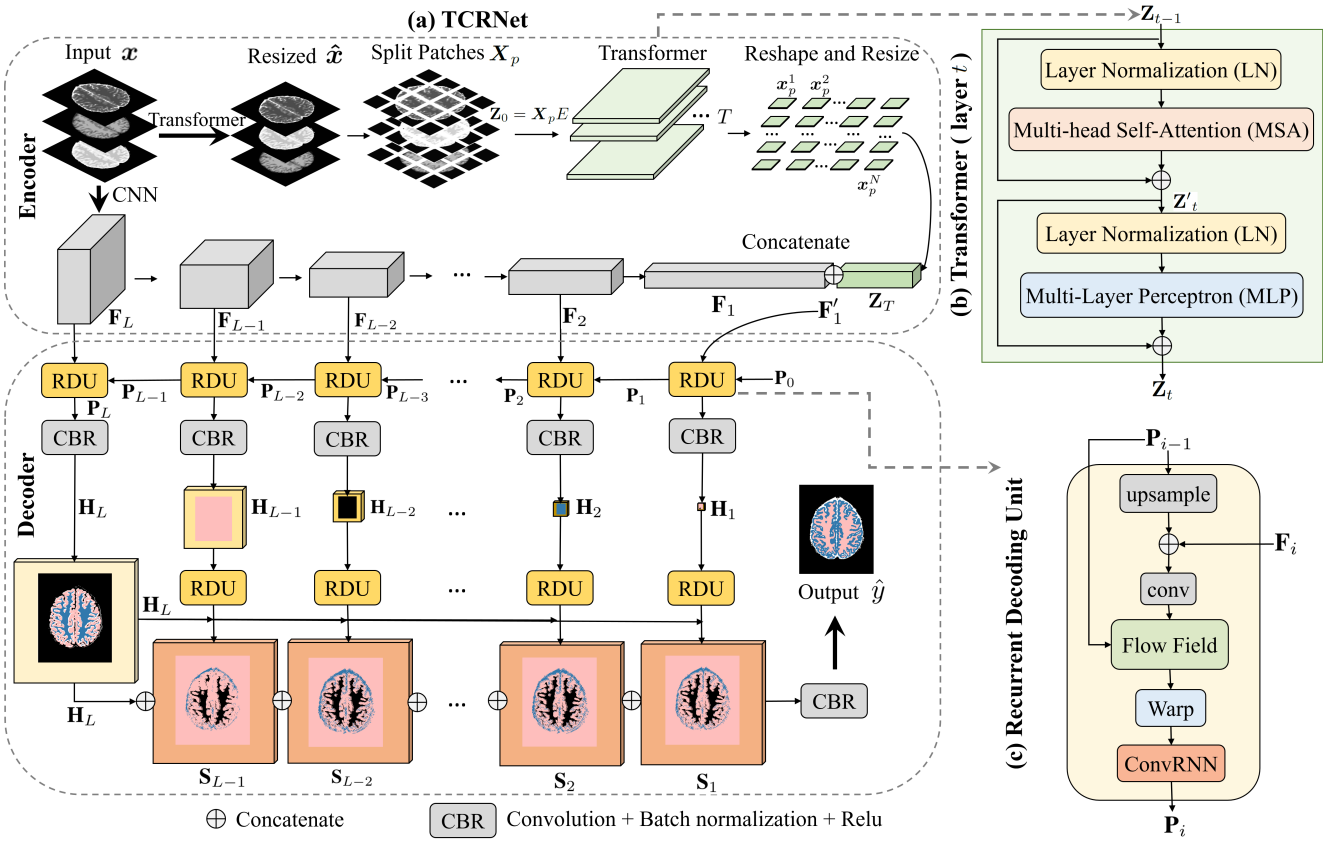

Fig. 1. The overview of the proposed method.

CNN to extract the global and local feature information, then we fuse the feature maps and get the prediction in the decoder.

2.1. Transformer and CNN in the Proposed Encoder

As the first row of the encoder in Fig. 1 (a), given an input image $\boldsymbol{x} \in \mathbb{R}^{C\times H\times W}$ (multi-modality inputs are concatenated at the channel for processing), to satisfy the input requirement of Transformer, $\boldsymbol{x}$ needs to be resized as $\hat{\boldsymbol{x}} \in \mathbb{R}^{C\times R\times R}$ and then be split into square patches $\boldsymbol{X}_p = \left[\boldsymbol{x}_p^1, \boldsymbol{x}_p^2, ..., \boldsymbol{x}_p^N\right] \in \mathbb{R}^{N\times(P^2\times C)}$, where C is the number of channels, $R \times R$ is the size of $\hat{\boldsymbol{x}}$, $P \times P$ is the size of $\boldsymbol{x}_p$, $N = \frac{R\times R}{P\times P}$ is the number of patches. Then we flatten $\boldsymbol{x}_p$ and map it to a D-dimensional patch embedding $\boldsymbol{X}_pE$ by the linear projection E. After generating $\boldsymbol{X}_pE$, we directly encode it by Transformer, instead of adding a class token to $\boldsymbol{X}_pE$, which is a process different from ViT. There are T alternating layers of multi-head self-attention (MSA) and multi-layer perceptron (MLP) blocks [8] in Transformer as shown in Fig. 1 (b). Formally, the output $\mathbf{Z}_T$ of the Transformer-based encoder is:

$$\mathbf{Z}_0 = \boldsymbol{X}_pE = \left[\boldsymbol{x}_p^1E, \boldsymbol{x}_p^2E, ..., \boldsymbol{x}_p^NE\right] \quad (1)$$

$$\mathbf{Z'}_t = \mathrm{MSA}(\mathrm{LN}(\mathbf{Z}_{t-1})) + \mathbf{Z}_{t-1} \quad (2)$$

$$\mathbf{Z}_t = \mathrm{MLP}(\mathrm{LN}(\mathbf{Z'}_t)) + \mathbf{Z'}_t \quad (3)$$

where $t = \{1, 2, ...T\}$ denotes the number of layers, LN$(\cdot)$ is the layer normalization and $\mathbf{Z'}_t$ is the layer t output of MSA.

Since Transformer is only good at capturing global information, CNN is used to complement the local information. For each input $\boldsymbol{x}$, as the second row of the encoder in Fig. 1 (a), we use a L-layer CNN-based encoder to extract its multi-scale feature maps $\{\mathbf{F}_1, \mathbf{F}_2, ..., \mathbf{F}_L\}$, where $\mathbf{F}_i \in \mathbb{R}^{c_i\times h_i\times w_i}$, $i = \{1, 2, ..., L\}$, c_i, h_i and w_i respectively denote the number of channels, height and width of the i-th feature map.

In order to concatenate $\mathbf{Z}_T$ from Transformer and $\mathbf{F}_1$ from CNN for decoding, due to the different feature size, we resize $\mathbf{Z}_T \in \mathbb{R}^{N\times P\times P}$ as $\mathbf{Z}_T \in \mathbb{R}^{N\times h_1\times w_1}$, so the concatenation of $\mathbf{Z}_T$ and $\mathbf{F}_1$ is $\mathbf{F}'_1 \in \mathbb{R}^{(N+c_1)\times h_1\times w_1}$.

2.2. Recurrent Decoding Unit in the Proposed Decoder

Considering the ability of RNN to mine the temporal and context information of the features from the encoder, we combine convRNN with flow field to construct a new recurrent decoding unit (RDU) as is shown in Fig. 1 (c). Starting with $\mathbf{F}'_1$, we exploit RDU to get a fined feature map $\mathbf{P}_1$. Then layer by layer, $\mathbf{P}_i$ can be integrated by $\mathbf{P}_{i-1}$ and $\mathbf{F}_i$ as:

$$\mathbf{P}_1 = \mathrm{RDU}(\mathbf{P}_0,\ \mathbf{F}'_1;\ \phi),\quad \mathbf{P}_i = \mathrm{RDU}(\mathbf{P}_{i-1},\ \mathbf{F}_i;\ \phi) \quad (4)$$

where $i = \{2, 3, ..., L\}$, $\mathbf{P}_i \in \mathbb{R}^{c_i \times h_i \times w_i}$, $\mathbf{P}_0$ is initialized as a zero tensor, and ϕ represents all the parameters in RDU. Following, Eq. (4) will be illustrated refer to Fig. 1 (c). We first concatenate $\mathbf{F}_i$ and the upsampled $\mathbf{P}_{i-1}$ (the size of $\mathbf{P}_{i-1}$ is the same as that of $\mathbf{F}_i$). Then, to make $\mathbf{P}_{i-1}$ higher-resolution, we perform a convolution operation on the concatenation to predict a flow field [14] $\Phi \in \mathbb{R}^{2 \times h_i \times w_i}$. For each pixel p of the image domain Ω, we warp $\mathbf{P}_{i-1}$ by $\Phi(p)$:

$$\text{Warp}(\mathbf{P}_{i-1}) = \mathbf{P}_{i-1}(p_x + \Phi(p)_x, p_y + \Phi(p)_y) \quad (5)$$

where $\forall p \in \Omega$, the subscript x and y are the coordinates of pixel p. Besides, the convRNN operation of RDU is:

$$\mathbf{P}_i = \sigma(\mathbf{W}_P \otimes \mathbf{P}_{i-1} + \mathbf{W}_F \otimes \mathbf{F}_i) \quad (6)$$

where $\sigma(.)$ denotes the ReLU activation function, $\otimes$ is the convolution operation, $\mathbf{W}_P$ and $\mathbf{W}_F$ are the weight matrices respectively for the hidden state $\mathbf{P}_{i-1}$ and the input $\mathbf{F}_i$.

In order to predict, we perform the CBR (convolution, batch normalization and rectified linear unit) on $\{\mathbf{P}_1, \mathbf{P}_2, ..., \mathbf{P}_L\}$ to get the score maps $\{\mathbf{H}_1, \mathbf{H}_2, ..., \mathbf{H}_L\}$, where $\mathbf{H}_i \in \mathbb{R}^{C \times h_i \times w_i}$. In this case, the number of channels is reduced to the number of segmentation classes. Instead of using $\mathbf{H}_L$ directly as the final prediction, similar to Eq. (4), we also fuse $\{\mathbf{H}_1, \mathbf{H}_2, ..., \mathbf{H}_L\}$ by RDU to obtain the finer segmentation:

$$\mathbf{S}_i = \text{RDU}(\mathbf{H}_i, \mathbf{H}_L; \phi), \ i = 1, 2, ..., L-1 \quad (7)$$

where $\mathbf{S}_i \in \mathbb{R}^{C \times h_L \times w_L}$ is the score map. Finally, we make full use of the global and local information in TCRNet and perform the CBR on concatenated $\{\mathbf{S}_1, \mathbf{S}_2, ..., \mathbf{S}_{L-1}, \mathbf{H}_L\}$ to get the prediction $\hat{y} \in \mathbb{R}^{C \times H \times W}$.

3. EXPERIMENTS

3.1. Implement Details

Datasets. As shown in Fig. 2, BrainWeb [15], MRBrainS [16] and Choledoch [17] datasets are used in experiments. The MRIs in BrainWeb and MRBrainS are divided into white matter, gray matter, cerebrospinal fluid and background; the hyperspectral images in Choledoch are divided into cancerous and normal areas. There are three sizes of images in BrainWeb: 181×181, 217×181 and 181×217; the images in MRBrainS are 240×240, and the images in Choledoch are 1280×1024. The number of images in BrainWeb, MRBrainS and Choledoch datasets is 399, 174 and 514 respectively, from which we randomly choose 60% for training and 40% for testing.

Settings. The loss function is cross-entropy loss in TCRNet. In this paper, we implement all the experiments based on PyTorch 1.6.0 on Intel® Xeon® Gold 6230 CPU @ 2.10GHz (the total number of CPU cores is 40). The learning rate for Adam optimizer is initialized as 0.0006 and the weight decay is 0.0005 [13]. In addition, we set the size of batch as 1 and the total iteration as 30000. By default, for the input images of Transformer, $R = 256$, $N = 256$, and $P = 16$. Please see section 3.3 for selection strategies. The evaluation metrics are pixel accuracy (PA) and dice similarity coefficient (DSC).

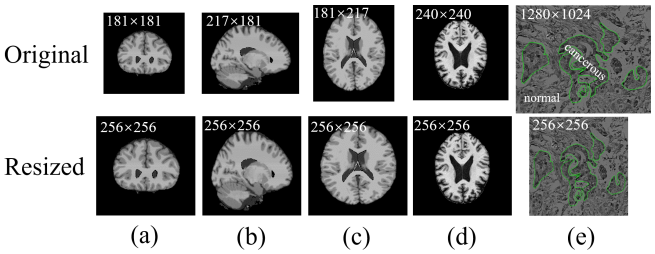

Fig. 2. The sample and resolution of images in BrainWeb: (a) (b) (c); MRBrainS: (d); Choledoch: (e).

Comparative Methods. There are eleven methods compared with TCRNet. Specifically, FCN-8s [18], U-Net [1], SegNet [19], U-Net 3+ [3], DFM [20] and SFNet [14] are CNN-based methods; LSTM-MA [12], BiLSTM-MA [12] and CRDN [13] take CNN as the encoder and RNN as the decoder; ViT [8] and TransUNet [9] take Transformer as the encoder. Refer to [9], we use naive upsampling as the decoder in ViT, and U-Net as the decoder in TransUNet.

3.2. Ablation Study

We investigate the importance of different components by ablation experiments. TCRNet without Transformer (#1) do not use Transformer to extract feature maps; TCRNet without CNN (#2) degrades to ViT because RDU needs the sequence of feature maps from CNN; TCRNet without RNN (#3) means that we perform one RDU and CBR on the concatenated feature of encoder to get prediction. It can be seen from Fig. 3 that TCRNet (#4) combines the merits of CNN, Transformer and RNN and shows the highest segmentation precision.

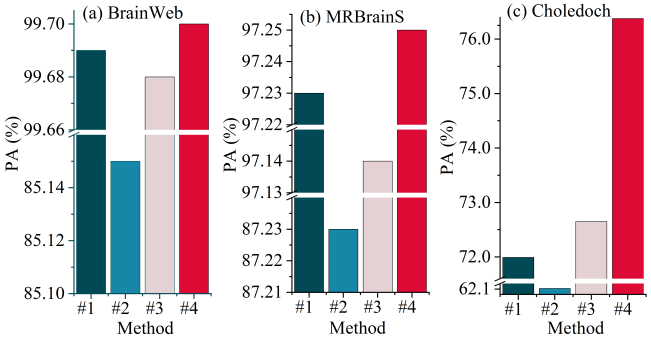

Fig. 3. Experimental results of ablation study.

3.3. Study on Diverse Image Resolution

To choose an appropriate image resolution, we explore the influence of diverse image resolution on segmentation performance. The experiments are conducted on the image resolution of 128×128 ($P = 8$) and 256×256 ($P = 16$). Since Transformer-based methods are sensitive to image resolution,

Table 1. Segmentation results of diverse image resolution on BrainWeb, MRBrainS and Choledoch datasets.

Method	Publication	Resolution	Metrics on BrainWeb		Metrics on MRBrainS		Metrics on Choledoch	
			PA (%)	DSC (%)	PA (%)	DSC (%)	PA (%)	DSC (%)
ViT [8]	ICLR'2021	128×128	83.14	65.37	87.48	60.91	64.00	57.39
TransUNet [9]	CVPR'2021	128×128	89.84	80.05	90.65	71.31	66.65	60.10
TCRNet	-	128×128	98.00	95.99	96.50	88.76	**74.25**	**70.55**
TCRNet	-	only Transformer 128×128	**99.68**	**99.36**	**97.25**	**90.85**	65.56	54.93
ViT [8]	ICLR'2021	256×256	85.15	69.53	87.23	61.96	62.11	56.79
TransUNet [9]	CVPR'2021	256×256	94.40	88.97	92.54	76.72	69.57	61.11
TCRNet	-	256×256	99.00	98.01	96.77	89.34	**76.38**	**72.80**
TCRNet	-	only Transformer 256×256	**99.70**	**99.40**	**97.26**	**90.85**	64.04	60.66

Table 2. Segmentation results of comparison on BrainWeb, MRBrainS and Choledoch datasets.

Method	Publication	Metrics on BrainWeb		Metrics on MRBrainS		Metrics on Choledoch	
		PA (%)	DSC (%)	PA (%)	DSC (%)	PA (%)	DSC (%)
FCN-8s [18]	CVPR'2015	90.30	78.89	92.61	75.70	66.35	61.84
U-Net [1]	MICCAI'2015	93.26	80.89	91.83	74.02	65.57	62.10
SegNet [19]	TPAMI'2017	93.36	86.13	92.44	75.60	65.66	61.46
LSTM-MA [12]	ICIP'2019	99.19	98.27	96.32	87.00	-	-
BiLSTM-MA [12]	ICIP'2019	99.38	98.66	96.33	87.02	-	-
U-Net 3+ [3]	ICASSP'2020	99.05	98.12	96.73	89.19	71.09	67.36
DFM [20]	MICCAI'2020	99.33	98.62	96.70	89.17	64.53	59.58
SFNet [14]	ECCV'2020	99.48	99.02	97.02	90.07	69.02	64.20
CRDN [13]	AAAI'2020	99.67	99.34	97.18	90.68	72.50	66.83
ViT [8]	ICLR'2021	85.15	69.53	87.23	61.96	62.11	56.79
TransUNet [9]	CVPR'2021	94.40	88.97	92.54	76.72	69.57	61.11
TCRNet	-	**99.70**	**99.40**	**97.25**	**90.86**	**76.38**	**72.80**

the comparative methods are ViT and TransUNet. We use two cases to explore whether to resize the input images of CNN, where 'only Transformer' means not to resize. The results are shown in Table 1.

We analyze the image resolution from Table 1. (i) The larger image resolution, the better segmentation: the results of 256×256 are better than those of 128×128. (ii) The regular square images such as MRBrainS dataset benefit to the results. Contrarily, the results of irregular images in Choledoch dataset are affected to some extent. (iii) Resizing the input images of CNN may improve the results. The performance of 'only Transformer' is poor in Choledoch dataset because the size of the resized image (256×256 or 128×128) greatly differs from that of the original image (1280×1024). To sum up, the appropriate image resolution is the integer power of 2 that is larger and closest to the original image resolution.

3.4. Comparison with State-of-the-art Methods

In order to testify the performance of TCRNet, we compare it with the comparative methods on BrainWeb, MRBrainS and Choledoch datasets. To ensure the fairness of comparison, U-Net is the backbone for all methods requiring backbones, and convRNN is used in CRDN. The resolution of images is 256×256 in ViT, TransUNet and TCRNet, 320×320 in U-Net 3+ as required. The rest comparative methods do not perform the resizing operation. Table 2 displays the experimental results.

TCRNet achieves the superb performance over the comparative methods. To be specific, compared with CNN-based and RNN-based methods such as CRDN, TCRNet gains 0.06%, 0.18% and 5.97% of DSC on the three datasets. In addition, the performance of Transformer-based methods such as TransUNet is unsatisfactory due to image deformation, while TCRNet concatenates the features extracted by Transformer and CNN, which weakens the effect of image deformation and increases 10.43%, 14.14% and 11.69% of DSC compared with TransUNet on the three datasets.

4. CONCLUSION

In this paper, we propose a TCRNet that makes Transformer, CNN and RNN complement each other for medical image segmentation. There are two solutions to tackle the problems in CNN-based and Transformer-based methods: (i) We concatenate the output of Transformer and CNN to weaken the effect of image deformation. (ii) We use Transformer to capture global feature information, CNN to retain local feature information and RNN to refine these feature information. The effectiveness of TCRNet is verified by extensive experiments.

5. REFERENCES

[1] N. Navab, J. Hornegger, W. Wells, and A. Frangi, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 12–20, 2015.

[2] Z. Zhou, M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Transactions on Medical Imaging*, vol. 39, no. 6, pp. 1856–1867, 2020.

[3] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y. Chen, and J. Wu, "Unet 3+: A full-scale connected unet for medical image segmentation," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1055–1059, 2020.

[4] Y. Zhang, J. Wu, Y. Liu, Y. Chen, E. Wu, and X. Tang, "MI-UNet: Multi-Inputs UNet Incorporating Brain Parcellation for Stroke Lesion Segmentation from T1-Weighted Magnetic Resonance Images," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 2, pp. 526–535, 2021.

[5] A. Chartsias, G. Papanastasiou, C. Wang, S. Semple, D. Newby, R. Dharmakumar, and S. Tsaftaris, "Disentangle, Align and Fuse for Multimodal and Semi-Supervised Image Segmentation," *IEEE Transactions on Medical Imaging*, vol. 40, no. 3, pp. 781–792, 2021.

[6] P. Chatterjee, K. Sharma, and A. Chakrabarti, "A stochastic approach for automated brain MRI segmentation," *IET Image Processing*, vol. 15, no. 3, pp. 735–745, 2021.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5999–6009, 2017.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D.Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *International Conference on Learning Representations (ICLR)*, pp. 1–22, 2021.

[9] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. Yuille, and Y. Zhou, "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation," *Proceedings of the IEEE International Conference on Computer Vision (CVPR)*, pp. 1–13, 2021.

[10] Y. Jin, D. Han, and H. Ko, "TrSeg: Transformer for semantic segmentation," *Pattern Recognition Letters*, vol. 148, pp. 29–35, 2021.

[11] Z. Peng, W. Huang, S. Gu, L. Xie, Y. Wang, J. Jiao, and Q. Ye, "Conformer: Local Features Coupling Global Representations for Visual Recognition," *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1–13, 2021.

[12] K. Xie and Y. Wen, "LSTM-MA : A LSTM Method with Multi-Modality and Adjacency Constraint for Brain image Segmentation," *IEEE International Conference on Image Processing (ICIP)*, pp. 240–244, 2019.

[13] Y. Wen, K. Xie, and L. He, "Segmenting Medical MRI via Recurrent Decoding Cell," *The Thirty-Forth AAAI Conference on Artificial Intelligence (AAAI)*, pp. 12452–12459, 2020.

[14] X. Li, A. You, Z. Zhu, H. Zhao, M. Yang, K. Yang, and Y. Tong, "Semantic Flow for Fast and Accurate Scene Parsing," *European Conference on Computer Vision (ECCV)*, pp. 775–793, 2020.

[15] R. Kwan, A. Evans, and G. Pike, "An extensible MRI simulator for post-processing evaluation," in *Visualization in Biomedical Computing*, pp. 135–140, 1996.

[16] A. Mendrik, K. Vincken, H. Kuijf, M. Breeuwer, W. Bouvy, J. De Bresser, A. Alansary, M. De Bruijne, A. Carass, A. El-Baz, A. Jog, R. Katyal, A. Khan, F. Van Der Lijn, Q. Mahmood, R. Mukherjee, A. Van Opbroek, S. Paneri, S. Pereira, M. Persson, M. Rajchl, D. Sarikaya, Ö. Smedby, C. Silva, H. Vrooman, S. Vyas, C. Wang, L. Zhao, G. Biessels, and M. Viergever, "MRBrainS Challenge: Online Evaluation Framework for Brain Image Segmentation in 3T MRI Scans," *Computational Intelligence and Neuroscience*, pp. 1–16, 2015.

[17] Q. Zhang, Q. Li, G. Yu, L. Sun, M. Zhou, and J. Chu, "A Multidimensional Choledoch Database and Benchmarks for Cholangiocarcinoma Diagnosis," *IEEE Access*, vol. 7, pp. 149414–149421, 2019.

[18] Y. Zhang, Z. Qiu, T. Yao, D. Liu, and T. Mei, "Fully Convolutional Adaptation Networks for Semantic Segmentation," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6810–6818, 2015.

[19] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[20] F. Cheng, C. Chen, Y. Wang, H. Shi, Y. Cao, D. Tu, C. Zhang, and Y. Xu, "Learning Directional Feature Maps for Cardiac MRI Segmentation," *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 108–117, 2020.