Mathematical Biosciences and Engineering

MBE, 20(2): 3177–3190.
DOI: 10.3934/mbe.2023149
Received: 02 October 2022
Revised: 21 November 2022
Accepted: 23 November 2022
Published: 02 December 2022

http://www.aimspress.com/journal/MBE

Research article

Hierarchical volumetric transformer with comprehensive attention for medical image segmentation

Zhuang Zhang[1] and Wenjie Luo[1,2,*]

[1] School of Cybersecurity and Computer, Hebei University, Baoding 071002, China
[2] Laboratory of Intelligence Image and Text, Hebei University, Baoding 071002, China

* **Correspondence:** Email: lwj12111@hbu.edu.cn.

Abstract: Transformer is widely used in medical image segmentation tasks due to its powerful ability to model global dependencies. However, most of the existing transformer-based methods are two-dimensional networks, which are only suitable for processing two-dimensional slices and ignore the linguistic association between different slices of the original volume image blocks. To solve this problem, we propose a novel segmentation framework by deeply exploring the respective characteristic of convolution, comprehensive attention mechanism, and transformer, and assembling them hierarchically to fully exploit their complementary advantages. Specifically, we first propose a novel volumetric transformer block to help extract features serially in the encoder and restore the feature map resolution to the original level in parallel in the decoder. It can not only obtain the information of the plane, but also make full use of the correlation information between different slices. Then the local multi-channel attention block is proposed to adaptively enhance the effective features of the encoder branch at the channel level, while suppressing the invalid features. Finally, the global multi-scale attention block with deep supervision is introduced to adaptively extract valid information at different scale levels while filtering out useless information. Extensive experiments demonstrate that our proposed method achieves promising performance on multi-organ CT and cardiac MR image segmentation.

Keywords: medical image segmentation; double transformer; local multi-channel attention; global multi-scale attention; convolutional neural network; deep supervision

1. Introduction

Medical image segmentation accurately captures the shape and volume of target organs and tissues through pixel-level classification, resulting in clinically useful diagnosis, treatment and intervention information [1]. Traditional medical image segmentation methods usually only rely on the low-level property of pixel-level features. Therefore, it is difficult for them to achieve satisfactory segmentation performance in the case of low contrast [2]. In addition, methods based on deep convolutional neural networks (CNN) have been applied to many medical image segmentation tasks, such as automatic multi-organ segmentation for computed tomography (CT) [3,4], colorectal polyp segmentation for colonoscopy videos [5], and prostate segmentation for magnetic resonance (MR) [6]. The convolution operation is essentially based on the basic mathematical operation between pixel values, and many excellent works based on mathematical models have emerged in recent years [7–9].

UNet [10] is undoubtedly a classic encoder-decoder framework for medical image segmentation, where the encoder extracts deep features by successive downsampling, and then the decoder uses the encoder outputs to continuously upsample to the original resolution. However, it may exhibit limitations in modeling explicit global dependencies due to the restricted receptive field of convolutional network. To address this problem, some studies utilized 3D convolution [11], attention mechanism [12], dilated convolution [13], and dynamic convolution [14] to efficiently extract feature information at different levels and increase the receptive field of the network. However, these methods still have certain limitations in capturing long distance dependencies due to the inherent restriction of convolution.

In recent years, transformer [15] has demonstrated excellent capabilities in natural language processing and is effective in learning global information and training on large-scale data [16,17]. It has recently attracted extensive attention and research in the fields of computer vision [18,19] and medical image processing [20,21]. Vision Transformer (ViT) [22] divided the input images into a series of patches for encoding and utilizes pure transformer blocks to model the global information. Swin Transformer [23] achieved the most advanced performance in a variety of computer vision tasks by generating hierarchical feature representations through self-attention layers with sliding windows of linear complexity rather than general quadratic. Chen et al. [20] first explored the potential of transformer in the field of medical image segmentation by employing transformer as the powerful encoders. Then, there have been many hybrid structures combining CNN and transformer [24–29]. By adding transformer to CNN-based networks, they can flexibly extract global and local feature information and achieve excellent segmentation performance. However, most of them only aggregate at the same level, ignoring the problem of semantic inconsistency between high-level and low-level features. Moreover, they only pass up the features of each level in decoder stage, not take full advantage of the features of different scales globally.

In this paper, we propose a novel hierarchical double transformer with comprehensive attention for accurate volumetric medical image segmentation. The main contributions are as follows:

1) We propose a new 3D model that hierarchically integrates convolutions, transformers, and comprehensive attention mechanisms to fully exploit their respective roles. Specifically, we first use convolution to perform preliminary low-level feature extraction, then use transformer to extract high-level features and downsample to reduce the amount of data and computation, and finally use multiple attention mechanisms to help enhance effective contextual semantic information.

2) From the perspective of network structure, we first introduce a novel double transformer structure used to extract features in encoding and restore the original resolution of feature maps in decoding. Then the local multi-channel attention mechanism (LMCA) and the global multi-scale attention mechanism (GMSA) adaptively filter out the effective features of the coding branch at the channel and scale levels in turn, while suppressing or filtering out the invalid features

3) The proposed method is implemented on two challenging medical image segmentation tasks, including abdominal multi-organ CT and cardiac MR datasets. Compared with the other four methods, our method achieves the highest Dice coefficient and obtains the best performance.

The rest of this article is organized as follows. Section 2 illustrates the specific details of our method. Section 3 describes the implementation details of the experiment and the experimental results. Section 4 provides the detailed discussion of the proposed modules. Finally, the conclusion is presented in Section 5.

2. Materials and methods

2.1. Datasets and pre-processing

We conducted extensive experiments on Synapse multiorgan segmentation (Synapse) dataset [30] and automated cardiac diagnosis challenge (ACDC) dataset [31]. The Synapse dataset includes 30 patients with a total of 3779 axial abdominal CT scan slices. Each case has 8 abdominal organs including aorta, gallbladder, spleen, left kidney, right kidney, liver, pancreas, and stomach. According to reference [20], 18 cases are extracted to construct the training set, and the remaining 12 cases are used for testing. The ACDC dataset includes 3D MRI images of 100 patients. The labels for each case include left ventricle (LV), right ventricle (RV), and myocardium (MYO). According to the setting of reference [20], the training data, validation data, and test data in the experiment are divided in a ratio of 7:1:2. All data are normalized and simple data augmentation, e.g., random rotation and flipping.

2.2. Overall structure

The overall structure of proposed method is shown in Figure 1, it includes two stages of encoder at the left end and decoder at the right end. This encoder and decoder structure has been used in many studies and achieved desirable performance [32–35]. In our model, it first utilizes a shrinking layer consisting of multiple convolutional layers to extract shallow features instead of a straightforward pure transformer. Then, the feature maps are passed to the double encoder block consisting of volume transformer and downsampling block. The double encoder block performs flat feature mapping and sequence-to-sequence position encoding on the incoming feature map. This encoding mode has both long implicit dependencies and small computational and spatial complexity. The volume transformer based on self-attention mechanism can adaptively adjust the receptive field according to the input feature. Fine extraction of higher-level features is continuously performed by stacking multiple hierarchical dual-encoding blocks and down-sampling blocks. Next, stacking multiple hierarchical upsampling blocks, local multi-channel attention blocks, and double decoder blocks progressively perform multiple pixel-level feature reconstruction. Finally, the global multi-scale attention block (GMSA) integrates features at different scale levels and outputs segmentation results. Through LCMA to eliminate the semantic gap between the left and right branch features at the channel level and GMSA

to adaptively enhance the effective information at different scale levels, it has certain pertinence to the blurred area and can accurately identify and segment target organs.

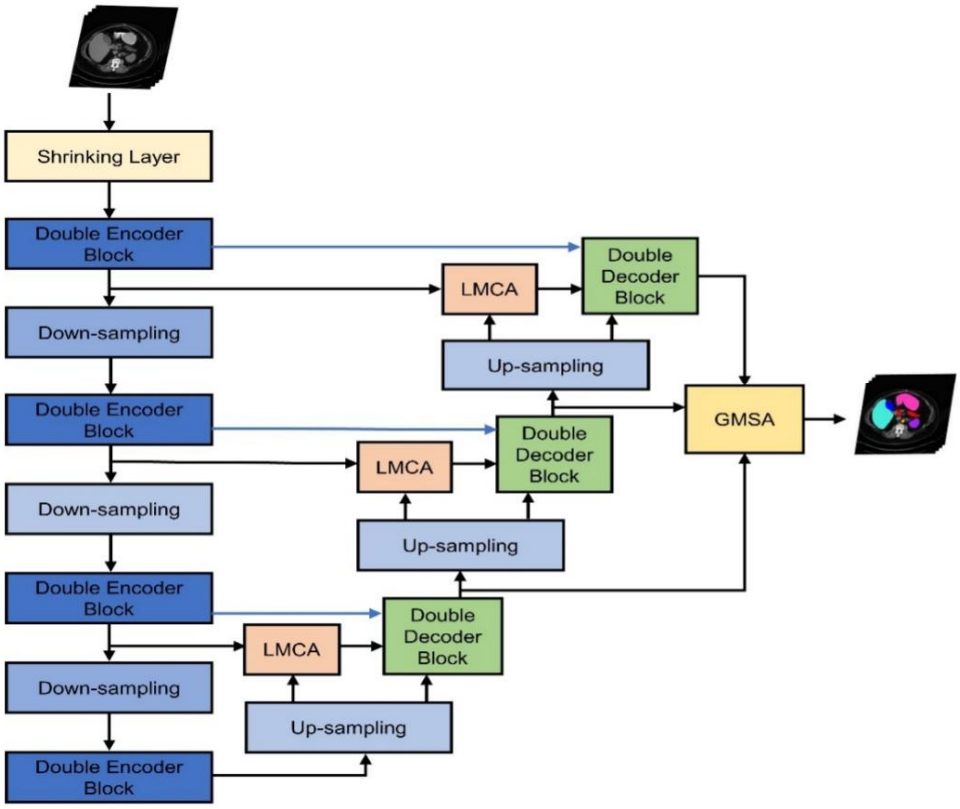

Figure 1. The overall structure of proposed method. The middle blue connecting lines represent the long-distance transfer of the key and value vectors of the transformer in the encoder stage.

2.3. The encoder stage

The encoder stage includes shrinking layer, double encoder block, and down-sampling. Unlike UNetr [36], we use convolution and transformer jointly to extract features. Specifically, the shrinking layer is composed by stacking four layers of convolution, batch normalization, and activation. It not only helps the extraction of shallow features but also relieves the computational burden of subsequent multiple transformers by reducing the volume size. The double encoder block consists of two layers of volume transformer which is shown in Figure 2. Each layer is mainly built on the volumetric multi-head self-attention (VMSA), followed by the Multi-Layer Perceptron (MLP) [15]. Specifically, given the input $X \in R^{C\times H\times W\times D}$ with the spatial resolution of $H \times W \times D$ and channels of C. It is first flattened and transposed into sequences with dimension of $C \times N$, where $N = H \times W \times D$. Then, it gets query ($Q \in R^{n\times N\times c}$), key ($K \in R^{n\times N\times c}$) and value ($V \in R^{n\times N\times c}$) through a linear layer, where n refers to the number of self-attention heads, $c = C / n$. After that, the following Eq (1) computes the attention function on query Q, key K, and value V.

$$Attention(Q,K,V) = (softmax(\frac{QK^T}{\sqrt{d_k}} + QL_B))V \tag{1}$$

Here, L_B refers to the learnable relative position encoding. As shown on the right side of Figure 2, different from the general relative position encoding, learnable position coding vectors L_v, L_h and L_d are first obtained simultaneously in the vertical, horizontal, and depth directions. Then we extend them to the same volumetric shape, add them together, and finally reshape them to match the shape of Q vector. After that, multiple self-attention heads are concatenated to get the output, which can be expressed as following Eq (2):

$$VMSA = Concat(head_1, head_2, \cdots, head_i) \tag{2}$$

where i refers to the number of head. Unlike the general structure, we do not add the original input directly to the result of self-attention. Instead, we introduce a learnable weight w to the original input, which is initialized to 1 and automatically learns the optimal weight value. Therefore, for an efficient transformer layer, X represent input, Z represent output. Its computational procedure can be summarized as follows:

$$X' = VMSA(Norm(X)) + wX \tag{3}$$

$$Z = MLP(Norm(X')) + X' \tag{4}$$

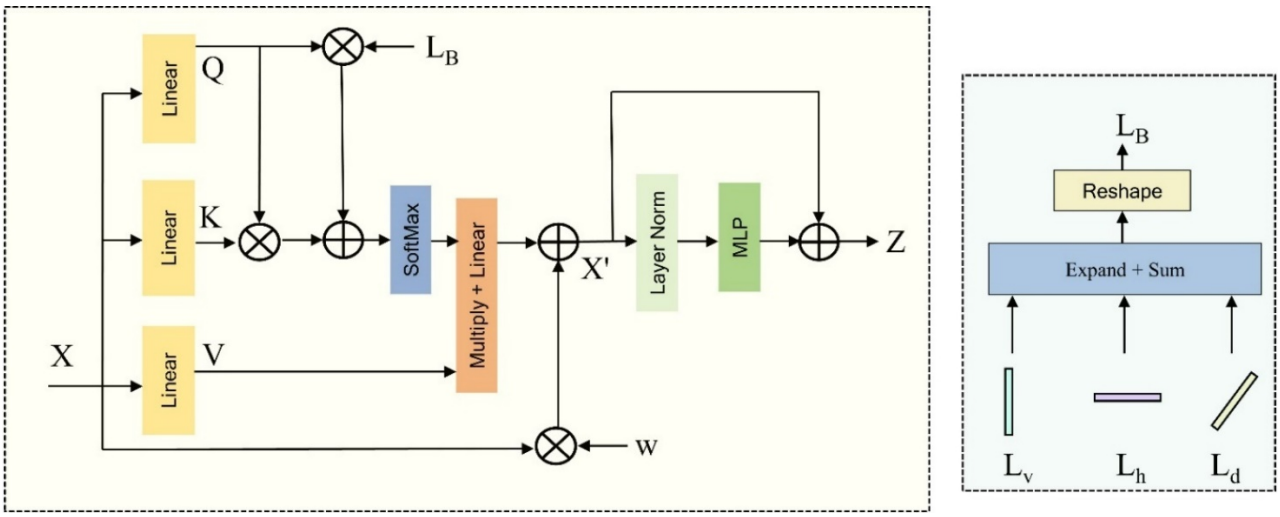

Figure 2. The structure of volumetric transformer in the encoder stage.

2.4. The decoder stage

The decoder part is shown on the right side of Figure 1 above, it mainly consists of upsampling, local multi-channel attention and double decoder blocks. Upsampling is achieved by 3D deconvolution. Inspired by various 2D attention mechanisms [37–39], our model also utilizes 3D attention

mechanisms. The local multi-channel attention structure is shown in Figure 3. It extracts effective information from the high-level features and low-level features propagated from the left end by stacking multi-channel average pooling, max pooling and multi-path dilated convolution (MPDC) blocks. The structure of MPDC is shown on the right of Figure 3. It mainly uses convolution blocks with different dilation rates of multipath to model feature information. The different dilation rates of each branch mean different receptive field sizes. By fusing features extracted from different receptive fields, more detailed global and local information can be obtained. Each convolution block of the module is stacked with 3-dimensional convolution with the kernel size of 1 or 3, group normalization and GELU activation functions. Specifically, the convolution blocks are first used to reduce the number of channels by one third of the original dimension, which can compress the information and reduce the computation. Then, the feature information is extracted simultaneously using three voids convolution blocks with cavity rates of 1, 2, and 4 respectively. Finally, they are splice together in channel dimension and depth-wise convolution is used to integrate the feature information. The information at different levels is then fused to help eliminate the semantic gap between the features at different levels at the encoding and decoding ends. Specifically, given the skip high level feature X_s and the low level feature X_l, the output of LMCA can be obtained by the following formulas:

$$Y_s = M(Ap(X_s)) + M(Mp(X_s)) \tag{5}$$

$$Y_l = M(Ap(X_l)) + M(Mp(X_l)) \tag{6}$$

$$LMCA = Sigmoid(C(Y_s, Y_l))X_s \tag{7}$$

where $M(.)$ refers to the output of the MPDC layer, $Ap(.)$ and $Mp(.)$ represent the output of the average pooling layer and the max pooling layer in turn, and C represents the concatenation at the channel level.

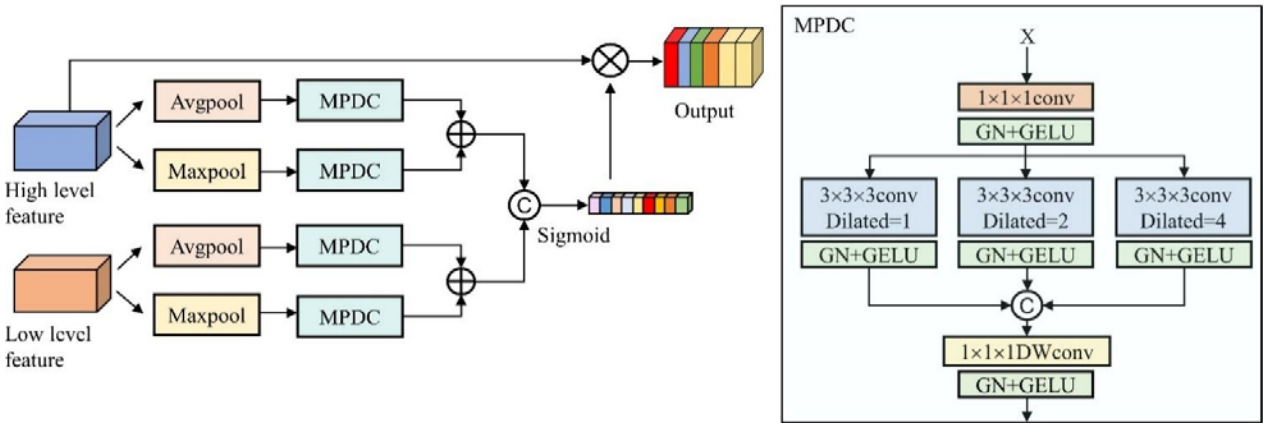

Figure 3. The structure of the LMCA block.

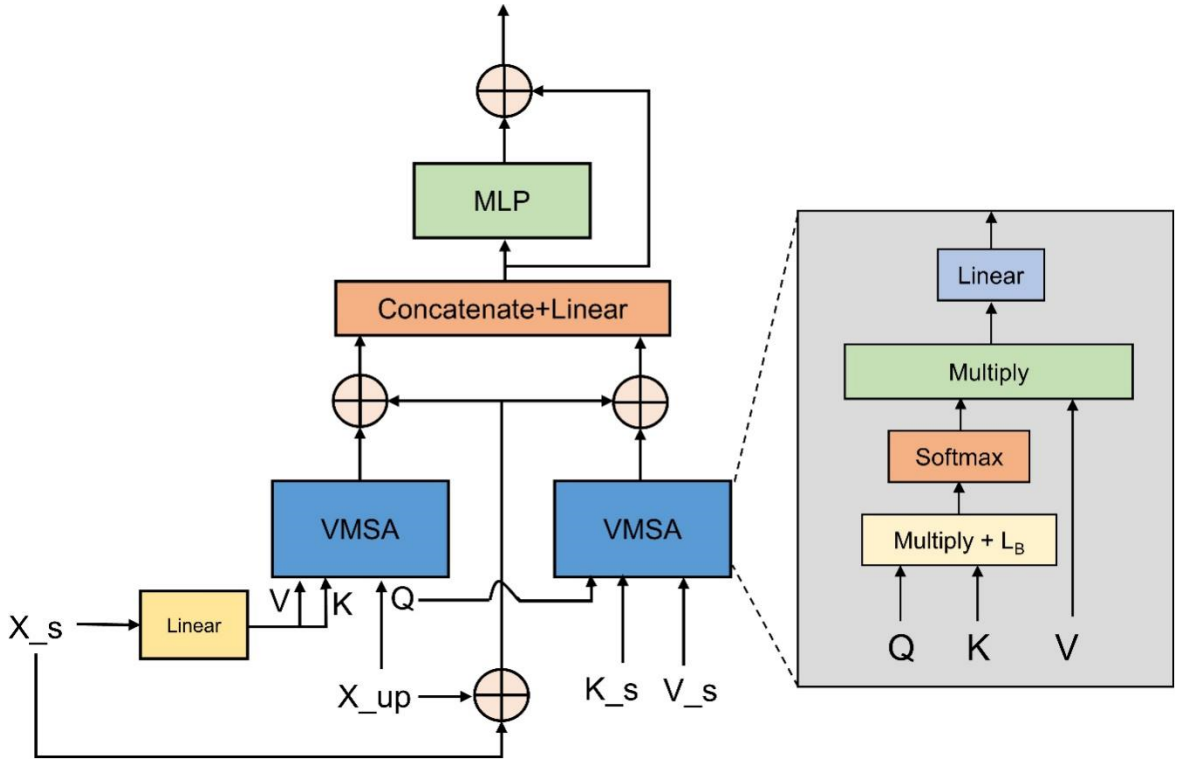

Figure 4. The structure of the volumetric transformer decoder block.

The double decoder block obtains high-level features by integrating the features of the lower branch, skip-connected features, skip-connected key, and value vectors. It mainly includes volumetric transformer decoder block. As shown in Figure 4, it consists of two parallel volumetric multi-head self-attention (VMSA) blocks and a multilayer perceptron layer. In the volumetric multi-head self-attention block on the left, the key and value vectors are computed by the features passed from encoder, while on the right, the key and value vectors are directly obtained by skip connection. The high-level semantic features can be obtained by splicing the attention maps of the left and right VMSA and sending them to the feedforward neural network. The specific calculation process is similar to the transformer in the encoder stage described in detail in Section 2.3.

Global scale attention integrates features at different scales. Its structure is shown in Figure 4, the low level two low-resolution features are first up-sampled to the same resolution as the high level feature, and stitched together in the channel dimension for channel-level and pixel-level information filtering. The weights at the channel level are implemented by max pooling and average pooling layers, while the pixel level is implemented by 3D convolutional layers. Through these two weights, the effective information can be adaptively enhanced and the invalid information can be suppressed. Specifically, given the high-resolution feature X_1 and two low-resolution features X_2, X_3, the output of GMSA can be obtained by the following formulas:

$$X = C(X_1, up(X_2), up(X_3)) \tag{8}$$

$$Y = Sigmoid(MLP(Ap(X)) + MLP(Mp(X)))X \tag{9}$$

$$GMSA = Sigmoid(\Theta(Y))Y \tag{10}$$

where C represents the concatenation at the channel level, $MLP(.)$ refers to the output of the multi-layer perceptual layer, $Ap(.)$ and $Mp(.)$ represent the output of the average pooling layer and the max pooling layer in turn, $\Theta(.)$ means two 3D convolutional layers.

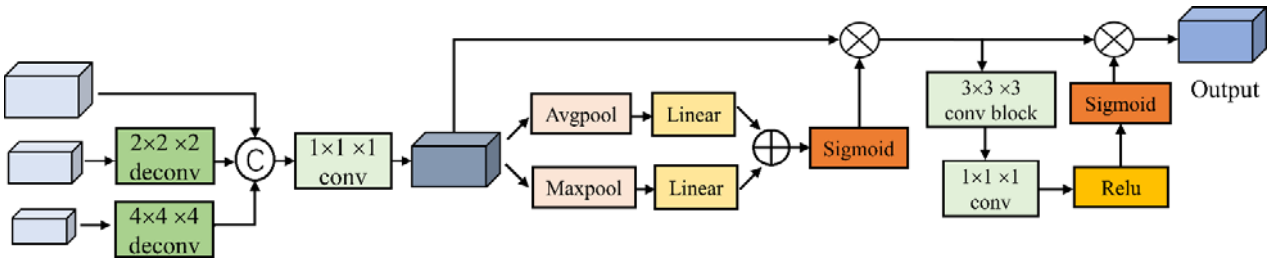

Figure 5. The structure of the GMSA block.

3. Experiment

3.1. Experiment details and evaluation metrics

We conducted experiments on the PyTorch platform with NVIDIA RTX A6000 and did not use any pre-trained weights to train the proposed network. For Synapse multi-organ dataset, we set the batch size of 2 and epoch of 1200. For ACDC dataset, the batch size is 2 and epoch is 1500. The stochastic gradient descent (SGD) optimizer is employed to train our model, where the initial learning rate is set to 0.01, the momentum is 0.99 and the weight decay is 3e-5. The evaluation metrics include dice coefficient (Dice) and 95% Hausdorff distance (95HD), they are calculated as follows:

$$Dice(X,Y) = \frac{2\left|X \bigcap Y\right|}{\left|X\right| + \left|Y\right|} \tag{11}$$

$$HD\left(X_s, Y_s\right) = \max_{x \in X_s}(\min_{y \in Y_s} ED(x,y)) \tag{12}$$

$$HD\left(X_s, Y_s\right) = \max(HD(X_s, Y_s), HD\left(Y_s, X_s\right)) \tag{13}$$

where X and Y represent the prediction and ground truth, ED is the Euclidean distance operator, while X_s and Y_s are the sets of surface points of the predicted and ground truth.

3.2. Experiments results

We evaluated the performance of the proposed method on multi-organ Synapse (CT) and cardiac ACDC (MR) datasets and compared with various state-of-the-art models including Unet [10], Transunet [20], Unetr [36], UTnet [40]. Table 1 reports the quantitative results on the Synapse dataset, where the evaluation metrics are Dice and 95HD. It can be seen that our method achieves the best

results on eight different organs and average dice similarity coefficient (DSC) of 88.91%, which shows our method significantly outperforms previous works. The results on the ACDC dataset are shown in Table 2, where the proposed method achieves the best results on the overall average results, RV and MYO, and is more than 3 percentage points higher than the other methods, while on LV only slightly lower than Transunet [20].

Table 1. Quantitative comparison of segmentation performance using different methods on Synapse dataset (Dice in %, 95HD in mm). The best performance is shown in bold.

Methods	Average Dice	Average 95HD	Aorta	Gallbladder	Kidney (L)	Kidney (R)	Liver	Pancreas	Spleen	Stomach
Unet [10]	76.85	39.70	89.07	69.72	77.77	68.60	93.43	53.98	86.67	75.58
Transunet [20]	77.48	31.69	87.23	63.16	81.87	77.02	94.08	55.86	85.08	75.62
UTnet [40]	77.65	26.73	86.68	62.50	83.58	75.33	94.83	59.02	85.62	73.68
Unetr [36]	79.56	22.97	89.99	60.56	85.66	84.80	94.46	59.25	87.81	73.99
Ours	88.91	10.63	92.73	80.24	87.61	87.37	96.95	82.73	94.10	89.56

Table 2. Quantitative comparison of segmentation performance using different methods on ACDC dataset (Dice Similarity Coefficient (DSC) in %). The best performance is shown in bold.

Methods	Average	RV	MYO	LV
Unet [10]	86.90	86.20	82.50	92.20
Transunet [20]	89.71	88.86	84.54	95.73
UTnet [40]	88.30	88.20	83.50	93.10
Unetr [36]	88.61	85.29	86.52	94.02
Ours	92.27	91.42	89.71	95.69

3.3. Visualization of segmentation results

We qualitatively compare the experimental results on Synapse data, and the visualization of its segmentation results is shown in Figure 6. It can be seen that the proposed method produces less class classification errors than other methods and is closer to the ground truth. Specifically, in the top row of picture in Figure 6, other methods misclassify the liver as the spleen or the right kidney misclassifies the left kidney. While in the middle row are errors for over- or under-segmentation of the stomach, it is clear that the proposed method is less prone to such errors.

The visualization of segmentation results on the ACDC dataset is shown in Figure 7. It can be seen that for the part of the right ventricle represented by the red area in the figure, various other methods have over-segmented errors, dividing some non-right ventricular parts into the right ventricle. Our method is significantly less prone to this error.

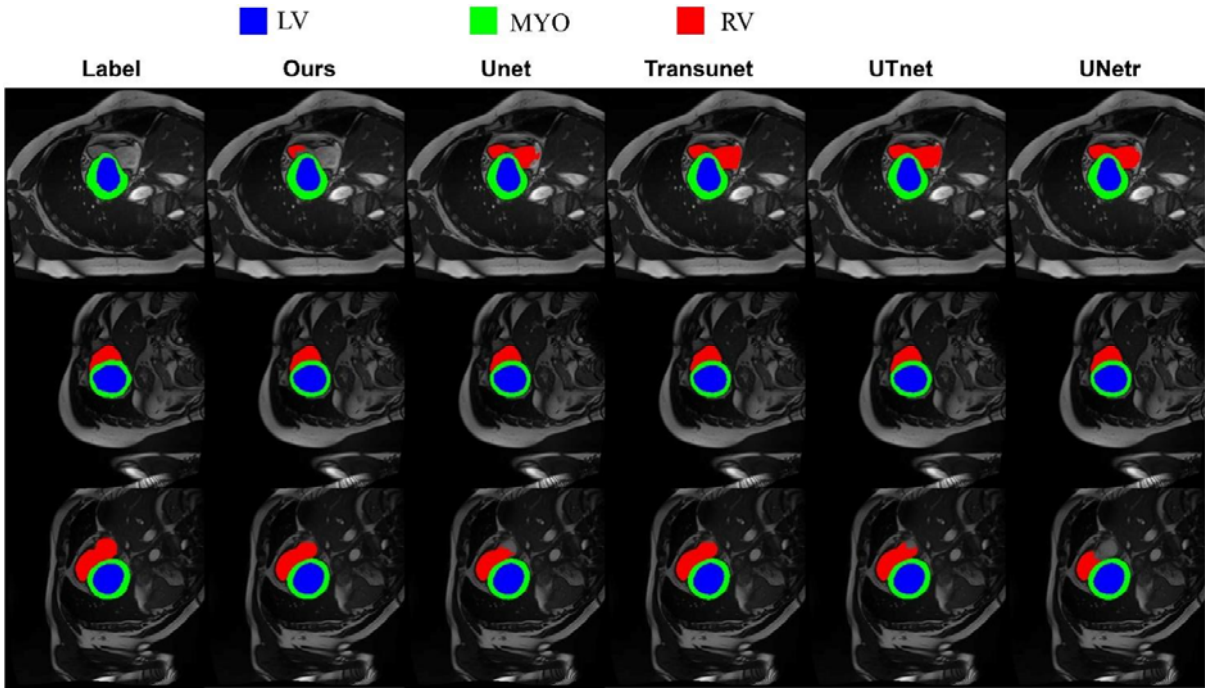

Figure 6. The segmentation results visualization in the Synapse dataset. The top is the color labeling of each organ. From left to right, ground truth and results of various methods are plotted on the original image. In particular, the areas where the differences are evident are clearly circled with red wireframes.

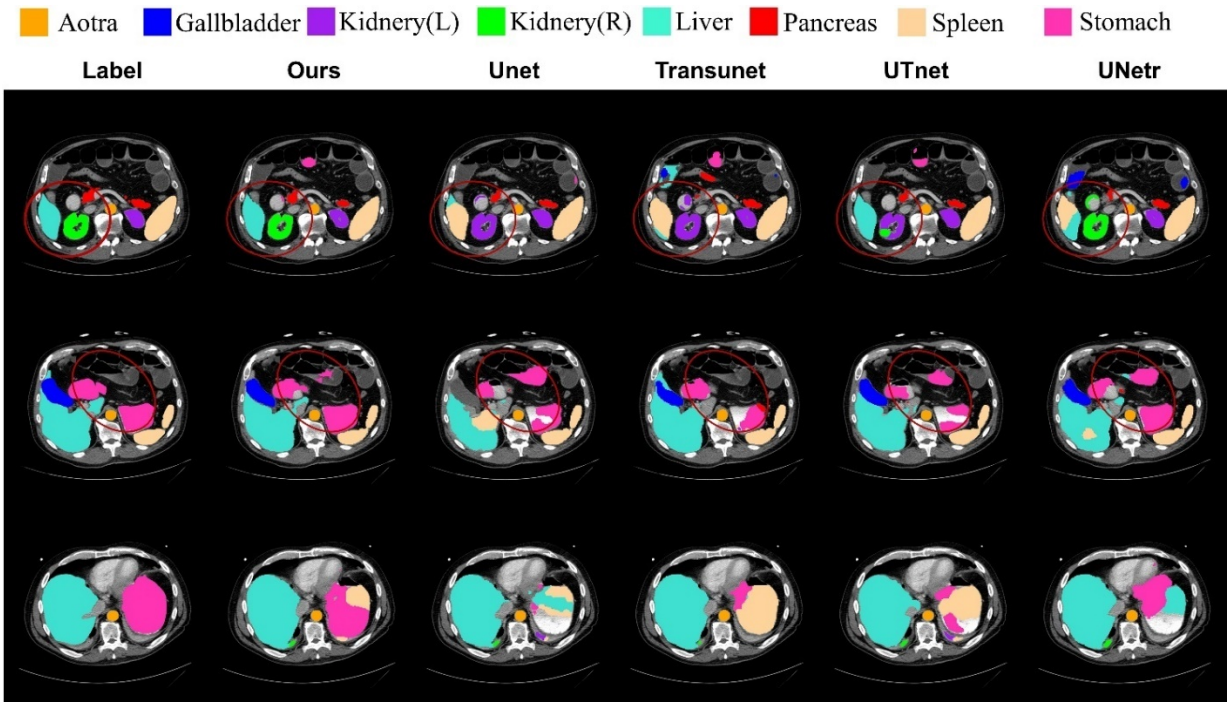

Figure 7. The segmentation results visualization in the ACDC dataset. The top is the color labeling of each organ. From left to right, ground truth and results of various methods are plotted on the original image.

4. Discussion

We discuss the proposed method on the Synapse dataset, performing detailed ablation experiments on each proposed component in the network. First, we denote the most primitive model with only volume transformers for encoding and decoding as VT. On the basis of VT, LMCA block, GMSA block, and volume transformer decoding (VTD) block are gradually added. Their detailed structures are discussed in Section 2 and correspond to Figures 3–5. Finally, deep supervision (DS) is performed on the feature maps of the three branches fed into the GMSA block and the final network output. Both feature maps use dice loss and cross entropy loss, and all their coefficients are set to 1. The results of the ablation experiments are shown in Table 3. The left end is the model with different components, and the right end is the corresponding dice coefficient on the Synapse dataset. It can be seen that the result of using only the volume transformer for encoding and decoding is 87.13%. Immediately after adding LMCA, GMSA, and VTD, it increased steadily by 0.28, 0.27, and 0.54 percentage points. Finally, the result reaches 88.51% by using deep supervision. These prove that each component can help guide accurate segmentation, especially VTD has a significant effect on the model.

Table 3. Quantitative comparison of segmentation performance using different components on Synapse dataset.

Method	Average Dice (%)
Ours (VT)	87.13
Ours (VT + LMCA)	87.41
Ours (VT + LMCA + GLMS)	87.74
Ours (VT + LMCA + GLMS + VTD)	88.28
Ours (VT + LMCA + GLMS + VTD + DS)	88.51

Table 4. Quantitative comparison of segmentation performance using different components on Synapse dataset.

Component	Existence	Dice (%)	95HD (mm)
LRPE	√	88.91	10.63
LRPE	×	88.34	11.57
MPDC	√	88.91	10.63
MPDC	×	88.45	10.98

We also conduct separate ablation experiments on the Synapse dataset for general transformers and proposed volume transformers with the learnable relative position encoding (LRPE) module. The experimental results are shown in Table 4. The volumetric transformer with LRPE and weight values resulted in a 0.57 percentage point improvement in Dice and a 0.96 mm increase in 95HD, respectively. In addition, we also conduct ablation experiments on the proposed MPDC module. As shown in Table 4, removing it means replacing it with a regular MLP layer in the network, and it brings a significant improvement in both Dice and 95HD.

5. Conclusions

In this paper, we propose a novel hierarchical volumetric transformer with comprehensive attention for accurate volumetric medical image segmentation. It first proposes a novel double transformer block to help extract features serially in the encoder and restore the feature map resolution to the original level in parallel in the decoder. Then the local multi-channel attention block is proposed to adaptively enhance the effective features of the encoder branch at the channel level, while suppressing the invalid features. Finally, the global multi-scale attention block with deep supervision is introduced to adaptively extract valid information at different scale levels while filtering out useless information. We demonstrate the remarkable effectiveness and robustness of the proposed method for medical image segmentation through extensive experiments.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, et al., A survey on deep learning in medical image analysis, *Med. Image Anal.*, **42** (2017), 60–88. https://doi.org/10.1016/j.media.2017.07.005
2. A. Sáez, C. Serrano, B. Acha, Model-based classification methods of global patterns in dermoscopic images, *IEEE Trans. Med. Imaging*, **33** (2014), 1137–1147. https://doi.org/10.1109/TMI.2014.2305769
3. Y. Zhao, H. Li, S. Wan, A. Sekuboyina, X. Hu, G. Tetteh, et al., Knowledge-aided convolutional neural network for small organ segmentation, *IEEE J. Biomed. Health Inf.*, **23** (2019), 1363–1373. https://doi.org/10.1109/JBHI.2019.2891526
4. X. Fang, P. Yan, Multi-organ segmentation over partially labeled datasets with multi-scale feature abstraction, *IEEE Trans. Med. Imaging*, **39** (2020), 3619–3629. https://doi.org/10.1109/TMI.2020.3001036
5. Y. Fang, C. Chen, Y. Yuan, K. Y. Tong, Selective feature aggregation network with area-boundary constraints for polyp segmentation, in *International Conference on Medical Image Computing and Computer Assisted Intervention*, (2019), 302–310. https://doi.org/10.1007/978-3-030-32239-7_34
6. F. Milletari, N. Navab, S. A. Ahmadi, V-net: Fully convolutional neural networks for volumetric medical image segmentation, in *2016 Fourth International Conference on 3D Vision (3DV)*, (2016), 565–571. https://doi.org/10.1109/3DV.2016.79
7. T. Jin, X. Yang, Monotonicity theorem for the uncertain fractional differential equation and application to uncertain financial market, *Math. Comput. Simul.*, **190** (2021), 203–221. https://doi.org/10.1016/j.matcom.2021.05.018
8. T. Jin, X. Yang, H. Xia, H. U. I. Ding, Reliability index and option pricing formulas of the first-hitting time model based on the uncertain fractional-order differential equation with Caputo type, *Fractals.*, **29** (2021), 21500122. https://doi.org/10.1142/S0218348X21500122
9. C. Tian, T. Jin, X. Yang, Q. Liu, Reliability analysis of the uncertain heat conduction model, *Comput. Math. Appl.*, **119** (2022), 131–140. https://doi.org/10.1016/j.camwa.2022.05.033.5

10. O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation, in *International Conference on Medical Image Computing and Computer-assisted Intervention*, (2015), 234–241. https://doi.org/10.1007/978-3-319-24574-4_28
11. Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, O. Ronneberger, 3D U-Net: Learning dense volumetric segmentation from sparse annotation, in *International Conference on Medical Image Computing and Computer-assisted Intervention*, (2016), 424–432. https://doi.org/10.1007/978-3-319-46723-8_49
12. S. Zhang, H. Fu, Y. Yan, Y. Zhang, Q. Wu, M. Yang, et al., Attention guided network for retinal image segmentation, in *Medical Image Computing and Computer Assisted Intervention*, (2019), 797–805. https://doi.org/10.1007/978-3-030-32239-7_88
13. W. Bo, Y. Lei, S. Tian, T. Wang, Y. Liu, P. Patel, et al., Deeply supervised 3D fully convolutional networks with group dilated convolution for automatic MRI prostate segmentation, *Med. Phys.*, **46** (2019), 1707–1718. https://doi.org/10.1002/mp.13416
14. J. Chen, X. Wang, Z. Guo, X. Zhang, J. Sun, Dynamic region-aware convolution, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2021), 8064–8073.
15. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al., Attention is all you need, in *Advances in Neural Information Processing Systems*, (2017), 1–11.
16. M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, A. Dosovitskiy, Do vision transformers see like convolutional neural networks, preprint, arXiv:2108.08810.
17. X. Zhai, A. Kolesnikov, N. Houlsby, L. Beyer, Scaling vision transformers, preprint, arXiv:2106.04560.
18. B. Cheng, A. G. Schwing, A. Kirillov, Per-pixel classification is not all you need for semantic segmentation, preprint, arXiv:2107.06278.
19. S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, et al., Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2021), 6881–6890.
20. J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, et al., TransUNet: Transformers make strong encoders for medical image segmentation, preprint, arXiv:2102.04306.
21. H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, et al., Swin-Unet: Unet-like pure transformer for medical image segmentation, preprint, arXiv:2105.05537.
22. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, et al., An image is worth 16x16 words: Transformers for image recognition at scale, preprint, arXiv:2010.11929.
23. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, et al., Swin transformer: Hierarchical vision transformer using shifted windows, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2021), 10012–10022.
24. J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, V. M. Patel, Medical transformer: Gated axial-attention for medical image segmentation, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, (2021), 36–46. https://doi.org/10.1007/978-3-030-87193-2_4
25. Y. Zhang, H. Liu, Q. Hu, Transfuse: Fusing transformers and cnns for medical image segmentation, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, (2021), 14–24. https://doi.org/10.1007/978-3-030-87193-2_2
26. W. Wang, C. Chen, M. Ding, H. Yu, S. Zha, J. Li, Transbts: Multimodal brain tumor segmentation using transformer, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, (2021), 109–119. https://doi.org/10.1007/978-3-030-87193-2_11

27. Y. Xie, J. Zhang, C. Shen, Y. Xia, Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, (2021), 171–180. https://doi.org/10.1007/978-3-030-87199-4_16
28. H. Wang, P. Cao, J. Wang, O. R. Zaiane, UCTransNet: Rethinking the skip connections in U-Net from a channel-wise perspective with transformer, preprint, arXiv:2109.04335.
29. H. Y. Zhou, J. Guo, Y. Zhang, L. Yu, L. Wang, Y. Yu, nnformer: Interleaved transformer for volumetric segmentation, preprint, arXiv:2109.03201.
30. B. Landman, Z. Xu, J. E. Igelsias, M. Styner, T. Langerak, A. Klein, Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge, *In Proc. MICCAI: Multi-Atlas Labeling Beyond Cranial Vault-Workshop Challenge*, **5** (2015), 12.
31. O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P. A. Heng, et al., Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved, *IEEE Trans. Med. Imaging*, **37** (2018), 2514–2525. https://doi.org/10.1109/TMI.2018.2837502
32. Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, 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*, (2018), 3–11. https://doi.org/10.1007/978-3-030-00889-5_1
33. H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, et al., Unet 3+: A full-scale connected unet for medical image segmentation, in *IEEE International Conference on Acoustics, Speech and Signal Processing*, (2020), 1055–1059. https://doi.org/10.1109/ICASSP40776.2020.9053405
34. Z. Gu, J. Cheng, H. Fu, K. Zhou, H. Hao, Y. Zhao, et al., CE-Net: Context encoder network for 2D medical image segmentation, *IEEE Trans. Med. Imaging*, **38** (2019), 2281–2292. https://doi.org/10.1109/TMI.2019.2903562
35. I. Fabian, P. F. Jaeger, S. A. A. Kohl, J. Petersen, K. H. Maier-Hein, nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation, *Nat. Methods*, **18** (2021), 203–211. https://doi.org/10.1038/s41592-020-01008-z
36. A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, et al., Unetr: Transformers for 3d medical image segmentation, in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, (2022), 574–584.
37. O. Ozan, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, et al., Attention u-net: Learning where to look for the pancreas, preprint, arXiv:1804.03999.
38. R. Gu, G. Wang, T. Song, R. Huang, M. Aertsen, J. Deprest, et al., CA-Net: Comprehensive attention convolutional neural networks for explainable medical image segmentation, *IEEE Trans. Med. Imaging*, **40** (2020), 699–711. https://doi.org/10.1109/TMI.2020.3035253
39. S. Woo, J. Park, J. Y. Lee, I. S. Kweon, Cbam: Convolutional block attention module, in *Proceedings of the European Conference on Computer Vision (ECCV)*, (2018), 3–19.
40. Y. Gao, M. Zhou, D. N. Metaxas, UTNet: A hybrid transformer architecture for medical image segmentation, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, (2021), 61–71. https://doi.org/10.1007/978-3-030-87199-4_6