Mathematical Biosciences and Engineering

MBE, 19(5): 5207–5222.
DOI: 10.3934/mbe.2022244
Received: 22 December 2021
Revised: 25 February 2022
Accepted: 14 March 2022
Published: 22 March 2022

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

Research article

TF-Unet：An automatic cardiac MRI image segmentation method

Zhenyin Fu[1,†], Jin Zhang[1,†], Ruyi Luo[2], Yutong Sun[3], Dongdong Deng[3,*] and Ling Xia[1,4,*]

[1] Key Laboratory for Biomedical Engineering of Ministry of Education, Institute of Biomedical Engineering, Zhejiang University, Hangzhou 310027, China
[2] Hangzhou Science and Technology Information Institute, Hangzhou 310026, China
[3] School of Biomedical Engineering, Dalian University of Technology, Dalian 116024, China
[4] Research Center for Healthcare Data Science, Zhejiang Lab, Hangzhou 310026, China

† These authors contributed equally to this study.

* **Correspondence:** Email: xialing@zju.edu.cn, dengdongdong@dlut.edu.cn.

Abstract: Personalized heart models are widely used to study the mechanisms of cardiac arrhythmias and have been used to guide clinical ablation of different types of arrhythmias in recent years. MRI images are now mostly used for model building. In cardiac modeling studies, the degree of segmentation of the heart image determines the success of subsequent 3D reconstructions. Therefore, a fully automated segmentation is needed. In this paper, we combine U-Net and Transformer as an alternative approach to perform powerful and fully automated segmentation of medical images. On the one hand, we use convolutional neural networks for feature extraction and spatial encoding of inputs to fully exploit the advantages of convolution in detail grasping; on the other hand, we use Transformer to add remote dependencies to high-level features and model features at different scales to fully exploit the advantages of Transformer. The results show that, the average dice coefficients for ACDC and Synapse datasets are 91.72 and 85.46%, respectively, and compared with Swin-Unet, the segmentation accuracy are improved by 1.72% for ACDC dataset and 6.33% for Synapse dataset.

Keywords: deep learning; neural networks; medical image segmentation; MRI

1. Introduction

Cardiac personalized modelling has been used for the non-invasive diagnosis and treatment of heart rhythm disorders, including risk classification of patients with heart attacks [1–3], prediction of the location of re-entry [4], and guidance for clinical ablation [5]. The key to the clinical application of heart models is the accurate creation of the personalized model, which is currently mostly segmented by experienced experts. Manual segmentation is subjective, irreproducible and time consuming, while model simulation takes a great deal of time and is time-critical in clinical practice, so how to minimize heart modeling time makes an automated segmentation method extremely important for the clinical application of personalized heart modelling.

Before the rise of deep learning, classical medical image segmentation algorithms such as region-based, grayscale-based, and edge-based algorithms were well established for medical images [6–9], while traditional machine learning techniques such as model-based methods (e.g., active contour and deformable models) and atlas-based methods (e.g., single-atlas and multi-atlas) have achieved good performance [10–13]. Nevertheless, both classical image segmentation algorithms and machine learning techniques usually require some prior knowledge or feature annotation processing to achieve better results. In contrast, deep learning-based algorithms do not rely on these operations; it automatically discovers and learns complex features from the data for target segmentation and detection. These features are often learned directly from the data through generic learning procedures and end-to-end methods. This allows deep learning-based algorithms to be easily applied to other domains. Deep learning-based segmentation algorithms are gradually surpassing previously advanced traditional methods and are gaining popularity in research, not only because of developments and advances in computer hardware such as graphics processing units (GPU) and tensor processing units (TPU), but also because of the increase in publicly available datasets and open source code. This trend can be observed in Figure 1, where the number of deep learning-based cardiac image segmentation papers has grown considerably in the last few years, especially after year 2018. We searched the web of science database for the keywords cardiac image segmentation and deep learning, and all types of articles were counted.

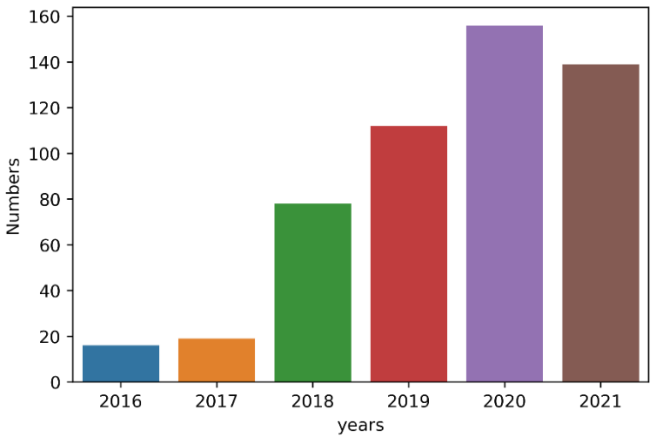

Figure 1. Overview of the number of papers published between 01 January 2016 and 01 December 2021 on deep learning-based methods for cardiac image segmentation.

Accurate localization and segmentation of medical images is a necessary prerequisite for diagnosis and treatment planning of cardiac diseases [14]. With the development of deep learning technology, various deep learning algorithms have been introduced into medical image processing and analysis with good results [15,16]. Convolutional neural networks (CNN) are one of the most common neural networks in medical image analysis, which are computationally fast and simple, and requiring no major adjustments to the network architecture [17]. CNN have been used with great success for medical image classification and segmentation, but a major drawback of this patch-based approach is that a separate network must be deployed for each patch at the time of inference, due to multiple overlapping in the image patches, which results in a large amount of redundancy and wasted resources. To solve this problem, fully convolutional networks [18] were created, which were designed to have an encoding-decoding structure that allows them to receive inputs of arbitrary size and produce outputs with the same size. However, this encoder-decoder structure also poses some limitations, such as the loss of some features, so there are many variants based on FCNs, but the most famous one is U-Net [19], which uses hopping connections to recover feature information in the down-sample paths to reduce the loss of spatial context information and thus obtain more accurate results. Subsequently U-Net gradually dominated in medical image processing, but it and its variants [20–22] also faced the lack of ability to build remotely correlated models. This is mainly due to the inherent limitations of the convolution operation [23].

On the other hand, the success of Transformer, which captures remote dependencies, has made possible the solution of the above problem in recent years. Transformer was designed for sequence modeling and transformation tasks, and it is known for its focus on modeling remote dependencies in data. Its great success in the language domain has motivated researchers to investigate its adaptability to computer vision, especially since it has achieved good results on some recent image classification and segmentation tasks [24–26]. ViT [25] first introduced transformer to computer vision tasks by segmenting an image into 16 non-overlapping patches, feeding them into standard transformer with positional embedding and comparing it with the CNN-based approach, ViT achieved a fairly good performance, which broke the monopoly of U-Net in computer vision. With the advent of VIT, more and more transformer-based image processing became popular, such as Swin-transformer [24] proposed a hierarchical transformer and a sliding window attention-based transformer, while the pyramidal visual transformer (PVT) [26] proposed a gradual shrinkage strategy to control the scale of feature maps and proposed a spatially reduced attention (SRA) layer to replace the traditional multiple head attention (MHA) layer in encoders, which were designed mainly to reduce the computational complexity. But these transformer-based networks have a limitation of unable extracting low-level features like convolutional operations [27], so some detailed features will be ignored.

To solve the above problem, we propose TF-Unet, a medical image segmentation framework that combines Transformer and U-Net. To fully utilize the advantages of both, we use two convolutional layers to learn high-resolution features and spatial location information in the learning feature phase, and use Transformer blocks to establish remote dependencies in the decoding phase. In terms of structure, inspired by the U-Net network structure, we divide the network into encoder-decoder blocks, and the self-attentive features of the coding blocks are combined with different high-resolution decoding features through hopping connections to reduce information loss. The results show that such a design allows our framework to maintain the advantages of both Convolution and Transformer, while facilitating the segmentation of medical images. Experimental

results show that our proposed hybrid network has better performance and robustness compared to previous methods based on pure convolution and pure transformer.

2. Materials and methods

2.1. Data description

The ACDC dataset: (1) Raw Nifti images of 100 patients were used as the training set, and clinical experts used the corresponding manual reference analysis of ED and ES time phases as segmentation criteria, where trabecular and papillary muscles were included in the ventricular blood pool; (2) raw Nifti images of another 50 patients were used as the test set, providing only basic patient information: height and weight, and ED and ES time phases. ACDC data were acquired using 1.5 T and 3.0 T MRI scanners with retrospective or prospective balanced steady-state free-feed sequences. The scan parameters were as follows: layer thickness of 5–8 mm, layer spacing of 5 mm, layer thickness and layer spacing combined were typically 5–10 mm, matrix size was 256 × 256, FOV was 300 × 330 mm^2, and one complete cardiac cycle consisted of 28–40 time phases.

The Synapse dataset: (1) Raw Nifti images of 30 patients were used as training set; (2) Raw Nifti images of another 20 patients were used as test set, these 50 scans were taken at the portal venography stage with different volumes (512 × 512 × 85–512 × 512 × 198) and fields of view (approximately 280 × 280 × 280–500 × 500 × 650 mm^3). The planar resolution varied from 0.54 × 0.54 to 0.98 × 0.98 mm^2, while the slice thickness ranged from 2.5 to 5.0 mm.

2.2. Methodology overview

The general architecture of TF-Unet is shown in Figure 2, which maintains a U-shape similar to that of U-net [11] and consists of two main branches, i.e., encoder and decoder. Specifically, the encoder includes the feature extraction block, the transformer block, and the down-sampling block. The decoder branch includes the transformer block, the up-sampling block and the deconvolution block that finally maps the output. And, to recover the image details in the prediction, we add residual connections [28] between the corresponding feature pyramids of the encoder and decoder in a symmetric manner.

2.2.1 Feature extraction block

The feature extraction block is mainly responsible for converting each input image I into a high-dimensional tensor $I \in R^{\frac{H}{4}\times\frac{W}{4}\times C}$, where H, W, C record the height, width, and sequence length of each input patch, respectively. Unlike Jieneng Chen et al. [24], who first flattened the input image directly and then preprocessed it in one dimension, we use a feature extraction layer which extracts low-level but high-resolution 3D features directly from the image and has more accurate spatial information at the pixel level.

We use two consecutive convolutional layers with a kernel size of 3 and step sizes of 2 and 1 and use LeakyReLU nonlinear activation functions and LayerNorm for each layer, which not only allows us to encode spatial information more accurately than the faceted position encoding used in the transformer, but also helps to reduce computational complexity while providing equally sized perceptual fields.

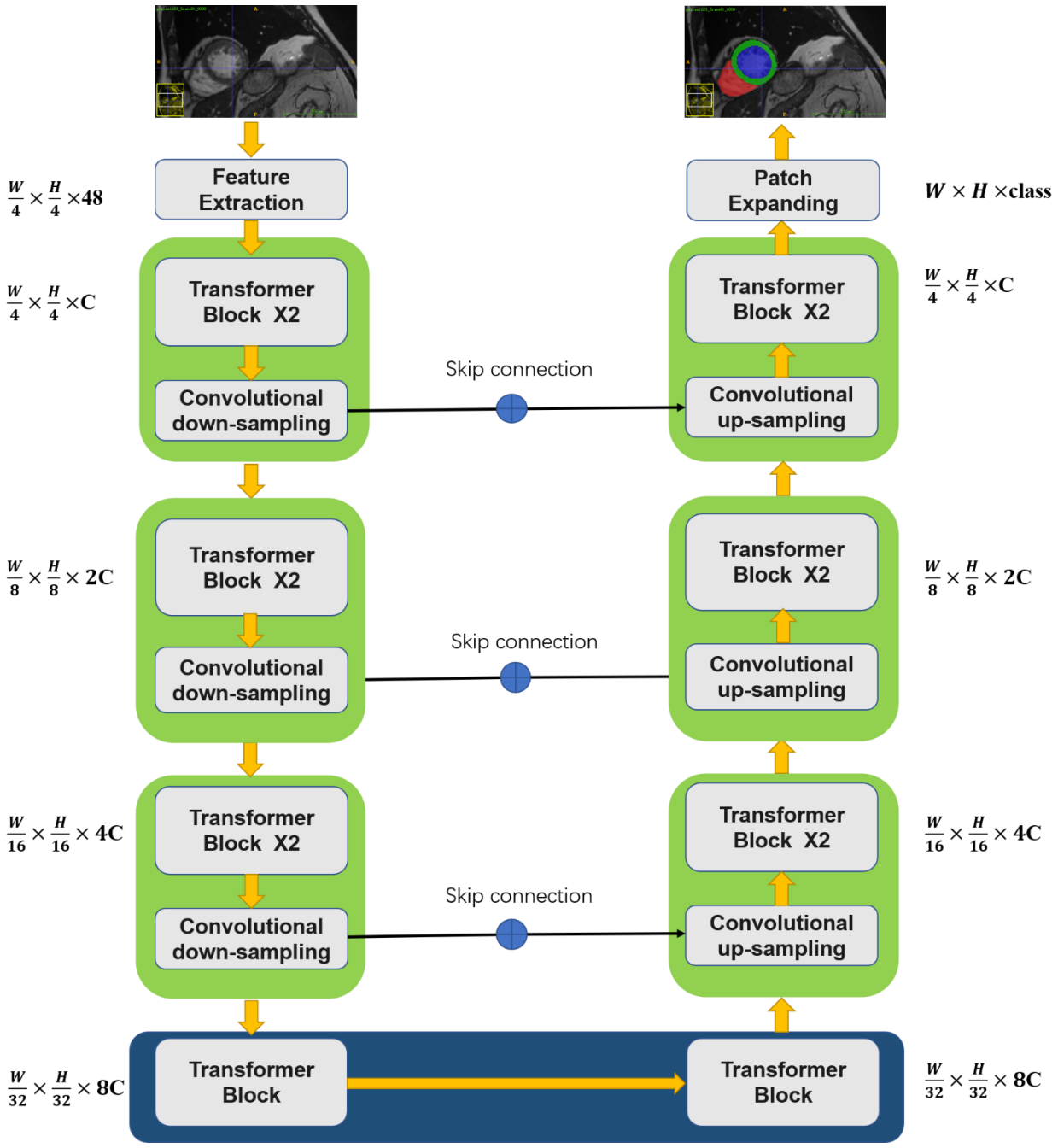

Figure 2. The overall architecture of TF-Unet, which is composed of encoder, decoder and skip connections. H, W, C represent the height, width and sequence length of each input patch, respectively.

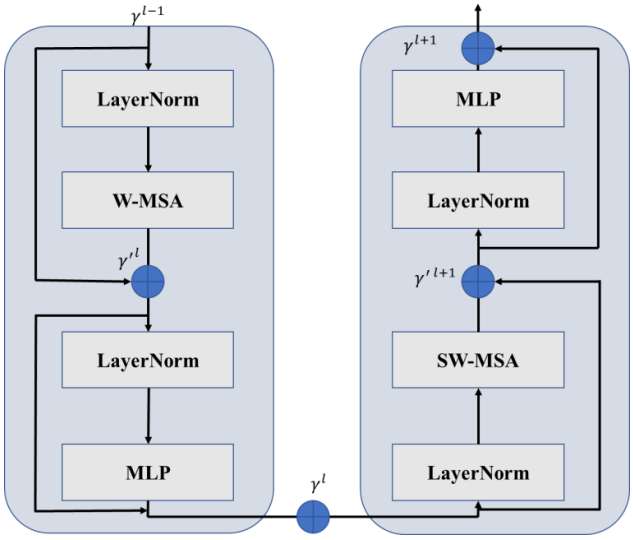

Figure 3. Two consecutive transformer blocks, the left is a fixed window, the right is a sliding window. Each transformer block is composed of LayerNorm layer, multi-head self-attention module and 2-layer MLP with LeakyReLU non-linearity.

2.2.2 Transformer block

After the feature extraction block, we pass the high-dimensional tensor I to the Transformer block in two consecutive layers. The ability of the transformer to establish remote dependencies is fully exploited to establish the connection between the high-resolution features extracted in the upper layer and the multi-scale features obtained by convolutional downsampling in the next layer. Unlike the traditional multi-headed self-attentive module, this paper uses the Swin-Transformer module [24], who is constructed based on a sliding window. Since the window-based self-attentive module lacks cross-window connections, this limits its modeling capabilities. In order to introduce cross-window connectivity while maintaining efficient computation of non-overlapping windows, Ze Liu et al. [24] proposed a sliding window partitioning approach. In Figure 3, two consecutive transformer modules are given. Each Swin-Transformer block consists of a LayerNorm (LN) layer, a multi-headed self-attentive module, a skip connection, and an MLP (Multilayer Perceptron) with a LeakyReLU nonlinearity. The window-based multi-headed self-attention (W-MSA) module and the sliding window-based multi-headed self-attention (SW-MSA) module are applied in the two consecutive Transformer blocks, respectively. Based on this window division mechanism, the consecutive sliding Transformer blocks can be represented as Eqs (1)–(4).

$$y'^{l} = W - MSA\left(LN\left(y^{l-1}\right)\right) + y^{l-1} \tag{1}$$

$$y^{l} = MLP\left(LN\left(y'^{l}\right)\right) + y'^{l} \tag{2}$$

$$y'^{l+1} = SW - MSA\left(LN\left(y^{l}\right)\right) + y^{l} \tag{3}$$

$$y^{l+1} = MLP\left(LN\left(y'^{l+1}\right)\right) + y'^{l+1} \tag{4}$$

where l is the index of the layer. W-MSA and SW-MSA denote the volume-based multi-headed self-attentive and its transfer version. where y'^{l} and y^{l} denotes the output of the W-MSA module and the MLP module of layer l, respectively. The computational complexity of SW-MSA on a volume of H × W × D patches is $4HWDC^2 + 2S_HS_WS_DHWDC$, however, the computational complexity of naïve multi-headed self-attention (MSA) is $4HWDC^2 + 2(HWD)^2C$. S_H, S_W, S_D represent the height, width and depth of the sliding window respectively. SW-MSA greatly reduces the computational complexity of MSA, so our proposed algorithm is more efficient. The sliding window segmentation approach introduces connections between adjacent non-overlapping windows in the previous layer and has been found to be effective in image classification, object detection and semantic segmentation [23].

In calculating the self-attention, we refer to Han Hu et al. [29,30] and add the relative position bias, and the specific formula for calculating the self-attention is as follows:

$$Attention(Q,K,V) = SoftMax\left(\frac{QK^T}{\sqrt{d}} + B\right)V \tag{5}$$

where Q, K, V represent the query matrix, key matrix and value matrix, respectively. d is generally taken as the dimension of Q or K. $B \in R^{(2H-1)\times(2W-1)}$ is the relative position encoding.

2.2.3 Convolutional down-sampling

Instead of completing the cascaded feature operations by using linear layers as in Swin-Unet [23], we directly use the convolution operation with stride size of 2. The reason for this is that the layered features generated by convolutional down-sampling help to model the target object at multiple scales. After such processing, the feature resolution is down-sampled by a factor of 2 and the feature dimension is increased to twice the original dimension.

2.2.4 Convolutional up-sampling

Corresponding to the Convolutional down-sampling, we also make changes in the up-sampling layer. We use stepwise deconvolution to up-sample the low-resolution feature map into a high-resolution feature map, i.e., by reconstructing the adjacent dimensional feature map into a higher-resolution feature map (2x up-sampling) and correspondingly reducing the feature dimension to half of the original dimension, and then by skip connecting, the features extracted from the encoder's down-sampling are combined with the decoder up-sampled features are merged. A deconvolution operation is also performed in the last patch extension block to produce the final result.

2.2.5 Skip connection

Similar to U-Net [19], skip connections are used to fuse multiscale features from the encoder with up-sample features from the decoder. We splice shallow and deep features together to reduce the loss of spatial information due to down-sample.

3. Results

To fairly compare the experimental results, we test three times on the ACDC dataset to take the average, and to verify the robustness of our algorithm, we do the same test on the Synapse dataset.

3.1. Experimental details

We ran all experiments based on Python 3.6, pytorch 1.8.1 and Ubutun 20.04. All training programs were executed on an NVIDIA 2080 GPU with 11 GB of RAM. The initial learning rate was set to 0.01, and we used the "poly" decay strategy [31] by default. As described in Eq (6):

$$lr = initial\_lr \times (1 - \frac{tem\_epoch}{max\_epoch})^{\gamma} \tag{6}$$

where max_epoch represents the total number of training generations, default 1000, tem_epoch represents the current training generations, γ is the hyperparameter, default take 0.9.

The default optimizer is stochastic gradient descent (SGD) and we set the momentum to 0.99. The weight decay is set to $3e^{-5}$. We use the weighted sum of the cross-entropy loss and the dice loss as the loss function. The training epochs is 1000 and each epoch contains 250 iterations.

3.2. Data pre-processing and data enhancement

All images in the same dataset are firstly resampled to the same target spacing and then cropped to the same size. Since there are not enough training samples, some data enhancement operations, such as rotation, scaling, Gaussian blur, Gaussian noise, brightness and contrast adjustment, are performed during the training process

3.3. Experimental result at ACDC

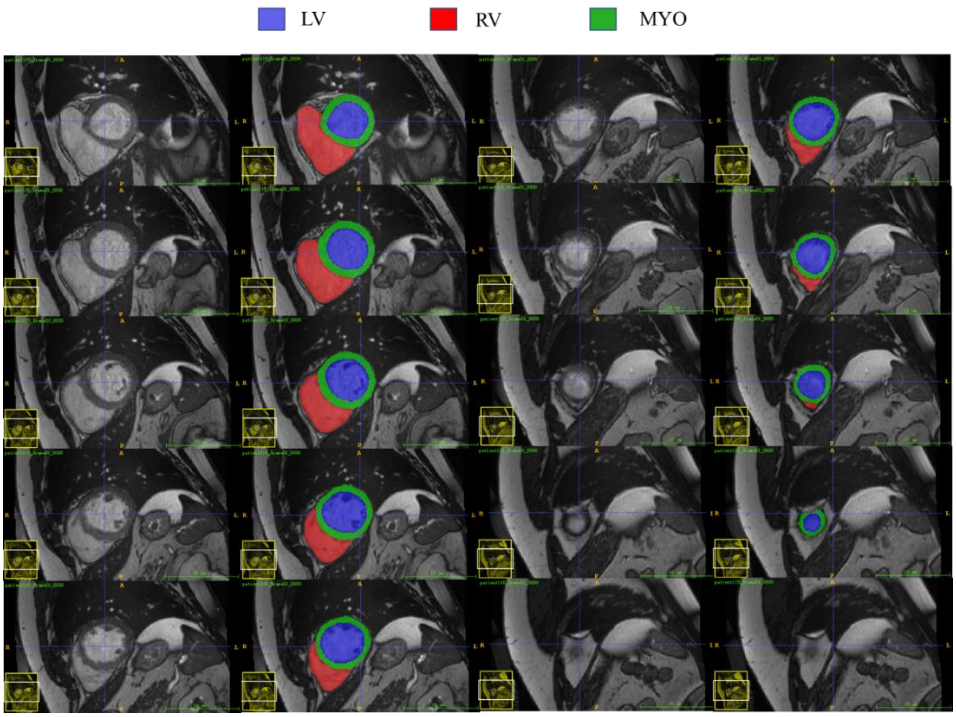

Figure 4. The patient's consecutive nine-layer cardiac MRI results, the first and third columns are the original images, sequentially from the base to the apex of the heart, and the second and fourth columns correspond to the segmentation results on the left, respectively.

In conducting experiments on the ACDC dataset, we designed two experimental scenarios; one is to make full use of the dataset, we use all 100 training data as the training set and 50 test data as the test set. The other is to quantitatively evaluate our results, we divide the 100 labeled training data into 70 training sets, 10 validation sets and 20 test sets. The real labels of the 20 cases used for testing were not put into the training.

Figures 4 and 5 show the results of the first and second scenario, respectively. We randomly selected several patients' results for visualization [32].

The results of the second scenario are as follows, Table 1 shows the quantitative calculations and comparisons of RV, MYO and LV using the dice coefficients, and Figure 5 shows the raw plots of several randomly selected patient data, ground truth and predicted results. Due to the random nature of data partitioning, the results of the other methods in Table I are taken from the results in the corresponding papers.

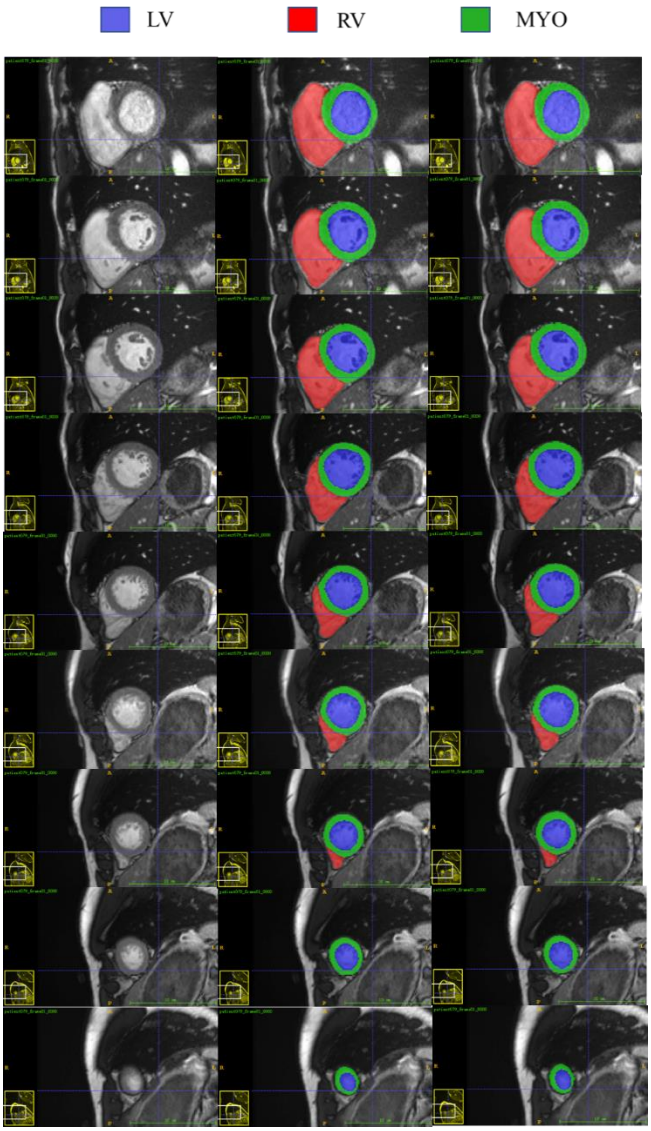

Figure 5. Cardiac MRI results in a patient with ground truth, the first columns are the original images, sequentially from the base to the apex of the heart, the second columns are the ground truth, and the third column are the segmentation results.

Table 1. Comparison on the ACDC MRI dataset (average dice score in %, dice score in % for each class).

Methods	DSC(avg)	RV	MYO	LV
R50-Unet [27]	87.55	87.10	80.36	94.92
R50-Attn Unet [33]	86.75	87.58	79.20	93.47
VIT [25]	81.45	81.46	70.71	92.18
R50-VIT [25]	87.57	86.07	81.88	94.75
TransUNet [27]	89.71	88.86	84.54	95.73
Swin-Unet [23]	90.00	88.55	85.62	**95.83**
Ours	**91.72**	**90.16**	**89.40**	95.60

3.4. Experimental result at Synapse

For Synapse data, we only use the second scenario in ACDC, we chose a part of labeled training set for testing, with training sample: validation sample: test sample = 14:7:9. We used the mean dice similarity coefficient (DSC) for eight abdominal organs, namely the aorta, gall bladder, spleen, left kidney, right kidney, liver, pancreas and stomach, to evaluate the model performance. Figure 6 shows the results of the different layers of patients from the Synapse dataset, with different colors representing different organs, as shown in the legend in Figure 6.

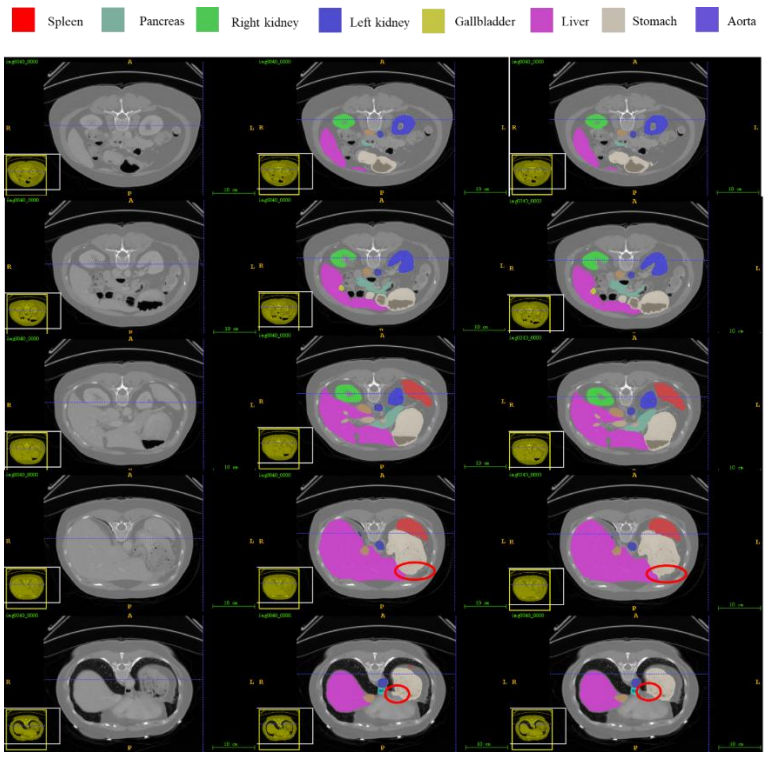

Figure 6. Results for one patient from the Synapse dataset, the first column is the original image, the second column is the ground truth, and the third column is the segmentation result. Due to the excessive number of scanned layers, we choose the intermediate layers to display, starting from layer 85 and displaying every ten layers at intervals until layer 125.

Table 2. Comparison on the Synapse multi-organ CT dataset (average dice score in %, dice score in % for each class).

Methods	DSC(avg)	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach
R50-Unet [27]	74.68	87.74	63.66	80.60	78.19	93.74	56.90	85.87	74.16
DualNorm-UNet [34]	80.37	86.52	55.51	88.64	86.29	95.64	55.91	94.62	79.80
VIT [25]	67.86	70.19	45.10	74.70	67.40	91.32	42.00	81.75	70.44
R50-VIT [25]	71.29	73.73	55.13	75.80	72.20	91.51	45.99	81.99	73.95
SQNet [35]	73.76	83.55	61.17	76.87	69.40	91.53	56.55	85.82	65.24
TransUNet [27]	77.48	87.23	63.16	81.87	77.02	94.08	55.86	85.08	75.62
Swin-Unet [23]	79.13	85.47	**66.53**	83.28	79.61	94.29	56.58	**90.66**	76.60
Ours	**85.46**	**87.45**	63.10	**92.44**	**93.05**	**96.21**	**79.06**	88.80	**83.57**

3.5. Ablation study

In this section, we introduce the importance of learning rate strategies. In order to verify the effect of different learning rate strategies on the results, we did controlled experiments with four functions, inv, multistep, poly, step, and the results of the four methods are shown in Figure 7.

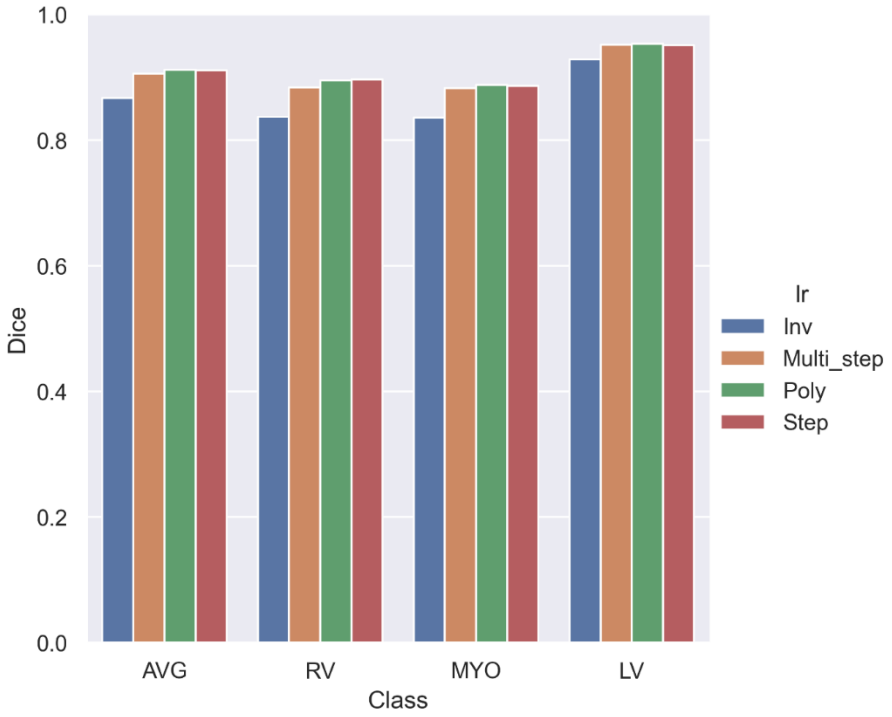

Figure 7. Results of four different learning rate strategies.

4. Discussion

In this section, we discuss in detail the experimental results obtained by our algorithm and explore the impact of different factors on the model performance, which we have compared on the ACDC and Synapse datasets, respectively. Specifically, we discuss the effects of different learning

rate strategies on network performance.

Analysis from a quantitative perspective. From Table 1, the best transformer-based model is Swin-Unet, which has an average dice coefficient of 90%. The best convolution-based model is R50-U-Net whose average dice coefficient is 87.55%, while our proposed TF-Unet is 1.72% higher than that of Swin-Unet and 4.17% higher than that of R50-U-Net. Considering that the current accuracy of these networks themselves is already very high, our proposed network improvement is still very effective, suggesting that our method can achieve better edge prediction. Analysis from a qualitative perspective. As can be seen in Figure 5, the middle represents the patient's true value and the rightmost represents our predicted value. By comparing layer by layer, the results obtained by our method are very close to the true value, and very good results are achieved even for the right ventricle, which is difficult to segment. In this work, we demonstrate that by combining Transformer with convolutional operations, better global and remote semantic information interactions can be learned, resulting in better segmentation results.

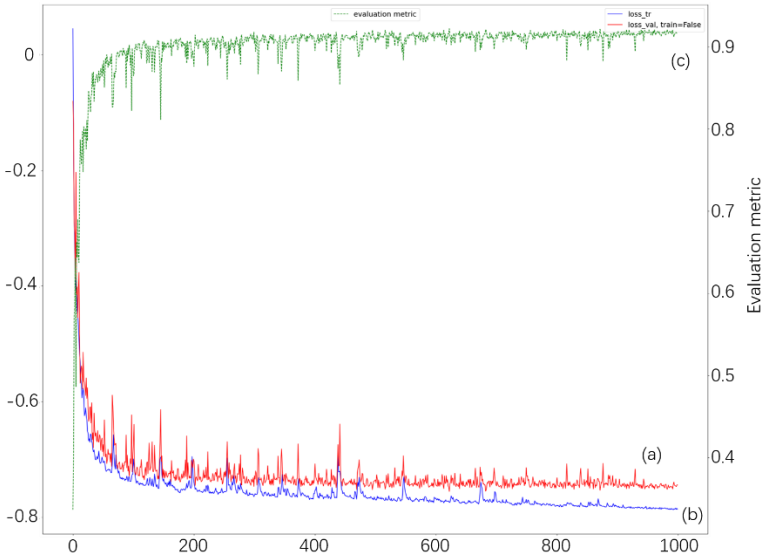

Figure 8. (a) The blue solid line represents training loss curves, (b) The red solid line represents validation loss curves, (c) The green dotted line represents dice score curves during validation.

It is well known that most of the deep learning networks cannot predict the results well for the test set without labeled values, but the network model based on TF-Unet can get good results. Observing Figure 4, we can conclude that the results obtained with our method are generally quite accurate for the layers other than the root tip layer. However, in the lower right corner of Figure 4, i.e., the apical layer, our method does not segment it. On the one hand, the apical layer has less segmentation in the training set, which makes it difficult for the network to learn features in this region; on the other hand, the true areas of both RV and LV in the apical layer are small and easily confused with surrounding vessels or tissues, leading to difficulties in segmentation.

Figure 8 summarizes the learning process of our proposed network, it was observed that the training loss and validation loss decrease with the increase of iterations and reach a stable state at about 200 generations without overfitting. And the dice coefficient of the validation set increases

with the number of iterations and reaches a steady state at 800 generations.

Quantitatively, as shown in Table 2, we performed experiments on Synapse and compared our TF-Unet with various transformer-based and Unet-based baselines. The main evaluation metric is the dice factor. the best performing transformer -based approach is Swin-Unet, which achieves an average score of 79.13. In contrast, DualNorm-UNet reports the best CNN-based results with an average of 80.37, slightly higher than Swin-Unet. our TF-Unet is able to outperform both Swin-Unet and DualNorm-UNet average performance by 6.33% and 5.09%, respectively, which is a considerable improvement on Synapse. Qualitatively, as can be seen in Figure 6, the middle column indicates the true value, and the rightmost column indicates the prediction result. For the segmentation of multiple organs, our proposed TF-Unet network still performs well, but there are some shortcomings for the stomach, as shown by the red boxes in the four lower right panels in Figure 6, one is that the prediction result is not smooth enough and there are many bursts, and the other is that it is difficult to segment to complex boundaries

Observing Figure 7 we can easily see that the results of all the functions are close except for the inv function. Through Figure 9 we speculate that this is because the learning rate of the inv function decreases too fast at the beginning of the iteration, and although it can speed up the search for the optimal solution, it is also easy to ignore the optimal solution and fall into the local optimal solution, leading to relatively poor results. The other three learning rates are all gradually decreasing, and although there is a big difference in the intermediate stages, the results do not differ much. These experiments show that the learning rate strategy has some influence on the experimental results, but it is generally enough to find the learning rate with the appropriate decreasing speed, and the different learning rate functions do not differ greatly.

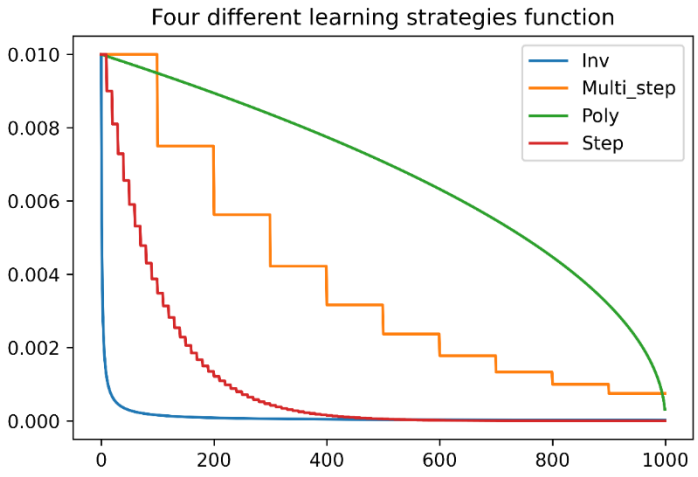

Figure 9. Four different learning rate functions.

5. Conclusions

In this paper, we propose a new medical image segmentation network TF-Unet. TF-Unet is built on the intertwined backbone of convolution and self-attention, which makes good use of the underlying features of CNN to build hierarchical object concepts at multiple scales through U-shaped

hybrid architectural design. In addition play Transformer's powerful self-attention mechanism that entangles long-term dependencies with convolutionally extracted features to capture the global context. Based on this hybrid structure, TF-Unet has made a great progress in previous Transformer-based segmentation methods. In the future, we hope that TF-Unet can replace manual segmentation operations for cardiac modeling, effectively improve the efficiency of personalized modeling, and accelerate the development of personalized cardiac models in clinical applications.

Acknowledgments

This study was supported by the Natural Science Foundation of China (NSFC) under grant number 62171408 and 81901841, the Key Research and Development Program of Zhejiang Province under grant number 2020C03016, the Major Scientific Project of Zhejiang Lab under grant number 2020ND8AD01, and Fundamental Research Funds for the Central Universities under grant number DUT21YG102.

Conflict of interest

The authors declare that there are no conflicts of interest.

References

1. E. Behradfar, A. Nygren, E. J. Vigmond, The role of Purkinje-myocardial coupling during ventricular arrhythmia: a modeling study, *PloS one*, **9** (2014), e88000. https://doi.org/10.1371/journal.pone.0088000
2. D. Deng, H. J. Arevalo, A. Prakosa, D. J. Callans, N. A. Trayanova, A feasibility study of arrhythmia risk prediction in patients with myocardial infarction and preserved ejection fraction, *Europace*, **18** (2016), iv60–iv66. https://doi.org/10.1093/europace/euw351
3. A. Lopez-Perez, R. Sebastian, M. Izquierdo, R. Ruiz, M. Bishop, J. M. Ferrero, Personalized cardiac computational models: From clinical data to simulation of infarct-related ventricular tachycardia, *Front. physiol.*, **10** (2019), 580. https://doi.org/10.3389/fphys.2019.00580
4. D. Deng, H. Arevalo, F. Pashakhanloo, A. Prakosa, H. Ashikaga, E. McVeigh, et al., Accuracy of prediction of infarct-related arrhythmic circuits from image-based models reconstructed from low and high resolution MRI, *Front. physiol.*, **6** (2015), 282. https://doi.org/10.3389/fphys.2015.00282
5. A. Prakosa, H. J. Arevalo, D. Deng, P. M. Boyle, P. P. Nikolov, H. Ashikaga, et al., Personalized virtual-heart technology for guiding the ablation of infarct-related ventricular tachycardia, *Nat. Biomed. Eng.*, **2** (2018), 732–740. https://doi.org/10.1038/s41551-018-0282-2
6. R. Pohle, K. D. Toennies, Segmentation of medical images using adaptive region growing, *Proc. SPIE*, **4322** (2002), 1337–1346. https://doi.org/10.1117/12.431013
7. C. Lee, S. Huh, T. A. Ketter, M. Unser, Unsupervised connectivity-based thresholding segmentation of midsagittal brain MR images, *Comput. Boil. Med.*, **28** (1998), 309–338. https://doi.org/10.1016/s0010-4825(98)00013-4
8. H. Y. Lee, N. C. Codella, M. D. Cham, J. W. Weinsaft, Y. Wang, Automatic left ventricle segmentation using iterative thresholding and an active contour model with adaptation on short-axis cardiac MRI, *IEEE Trans. Biomed. Eng.*, **57** (2010), 905–913. https://doi.org/10.1109/TBME.2009.2014545

9. S. Antunes, C. Colantoni, A. Palmisano, A. Esposito, S. Cerutti, G. Rizzo, Automatic right ventricle segmentation in ct images using a novel multi-scale edge detector approach, *Comput. Cardiol.*, (2013), 815–818.
10. P. Peng, K. Lekadir, A. Gooya, L. Shao, S. E. Petersen, A. F. Frangi, A review of heart chamber segmentation for structural and functional analysis using cardiac magnetic resonance imaging, *MAGMA*, **29** (2016), 155–195. https://doi.org/10.1007/s10334-015-0521-4
11. R. Hegadi, A. Kop, M. Hangarge, A survey on deformable model and its applications to medical imaging, *Int. J. Comput. Appl.*, (2010), 64–75.
12. V. Tavakoli, A. A. Amini, A survey of shaped-based registration and segmentation techniques for cardiac images, *Comput. Vision Image Understanding*, **117** (2013), 966–989. https://doi.org/10.1016/j.cviu.2012.11.017
13. D. Lesage, E. D. Angelini, I. Bloch, G. Funka-Lea, A review of 3D vessel lumen segmentation techniques: models, features and extraction schemes, *Med. Image Anal.*, **13** (2009), 819–845. https://doi.org/10.1016/j.media.2009.07.011
14. X. Liu, L. Yang, J. Chen, S. Yu, K. Li, Region-to-boundary deep learning model with multi-scale feature fusion for medical image segmentation, *Biomed. Signal Proc. Control*, **71** (2022), 103165. https://doi.org/10.1016/j.bspc.2021.103165
15. B. Pu, K. Li, S. Li, N. Zhu, Automatic fetal ultrasound standard plane recognition based on deep learning and IIoT, *IEEE Trans. Ind. Inf.*, **17** (2021), 7771–7780. https://doi.org/10.1109/TII.2021.3069470
16. J. Chen, K. Li, Z. Zhang, K. Li, P. S. Yu, A survey on applications of artificial intelligence in fighting against COVID-19, *ACM Comput. Surv.*, **54** (2021), 1–32. https://doi.org/10.1145/3465398
17. D. Ciresan, A. Giusti, L. Gambardella, J. Schmidhuber, Deep neural networks segment neuronal membranes in electron microscopy images, *Adv. Neural Inf. Proc. Syst.*, **2** (2012), 2843–2851. https://dl.acm.org/doi/10.5555/2999325.2999452
18. J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2015), 3431–3440. https://doi.org/10.1109/CVPR.2015.7298965
19. 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* (eds. N. Navab, et al.), Springer, Cham, **9351** (2015), 234–241. https://doi.org/10.1007/978-3-319-24574-4_28
20. F. Isensee, P. F. Jaeger, S. 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
21. 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 *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (2020), 1055–1059. https://doi.org/10.1109/ICASSP40776.2020.9053405
22. Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, J. Liang, Unet++: A nested u-net architecture for medical image segmentation, *Deep Learn. Med. Image Anal. Multimodal Learn. Clin. Decis. Support*, (2018), 3–11. https://doi.org/10.1007/978-3-030-00889-5_1
23. H. Cao, Y. Wang, J. Chen, et al., Swin-Unet: Unet-like pure transformer for medical image segmentation, preprint, arXiv:2105.05537.

24. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, et al., Swin transformer: Hierarchical vision transformer using shifted windows, preprint, arXiv:2103.14030.
25. A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., An image is worth 16x16 words: Transformers for image recognition at scale, preprint, arXiv:2010.11929.
26. W. Wang, E. Xie, X. Li, D. Weissenborn, X. Zhai, T. Unterthiner, et al., Pyramid vision transformer: A versatile backbone for dense prediction without convolutions, preprint, arXiv:2102.12122.
27. 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.
28. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2016), 770–778. https://doi.org/10.1109/CVPR.2016.90
29. H. Hu, J. Gu, Z. Zhang, J. Dai; Y. Wei, Relation networks for object detection, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2018), 3588–3597. https://doi.org/10.1109/CVPR.2018.00378
30. H. Hu, Z. Zhang, Z. Xie, S. Lin, Local relation networks for image recognition, in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2019), 3464–3473. https://doi.org/10.1109/ICCV.2019.00356
31. L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille, Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, *IEEE Trans. Pattern Anal. Mach. Intel.*, **40** (2017), 834–848. https://doi.org/10.1109/TPAMI.2017.2699184
32. P. A. Yushkevich, J. Piven, H. C. Hazlett, R. G. Smith, S. Ho, J. C. Gee, et al., User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability, *Neuroimage*, **31** (2006), 1116–1128. https://doi.org/10.1016/j.neuroimage.2006.01.015
33. J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, et al., Attention gated networks: Learning to leverage salient regions in medical images, *Med. Image Anal.*, **53** (2019), 197–207. https://doi.org/10.1016/j.media.2019.01.012
34. J. Xiao, L. Yu, L. Xing, A. Yuille, DualNorm-UNet: Incorporating global and local statistics for robust medical image segmentation, preprint, arXiv:2103.15858.
35. M. Treml, J. Arjona-Medina, T. Entertainer, R. Durgesh, F. Friedmann, P. Schuberth, et al., Speeding up semantic segmentation for autonomous driving, 2016.

AIMS Press