Two-Stage Segmentation Network with Feature Aggregation and Multi-Level Attention Mechanism for Multi-Modality Heart Images

Yuhui Song[b], Xiuquan Du[a,b], Yanping Zhang[a,b] and Shuo Li[c]

[a]Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, Anhui University
[b]School of Computer Science and Technology, Anhui University
[c]Department of Medical Imaging, Western University, London, ON N6A 3K7, Canada

ARTICLE INFO

Keywords:
Cardiac substructures
Segmentation
Feature aggregationn
Attention mechanism

ABSTRACT

Accurate segmentation of cardiac substructures in multi-modality heart images is an important prerequisite for the diagnosis and treatment of cardiovascular diseases. However, the segmentation of cardiac images remains a challenging task due to (1) the interference of multiple targets, (2) the imbalance of sample size. Therefore, in this paper, we propose a novel two-stage segmentation network with feature aggregation and multi-level attention mechanism (TSFM-Net) to comprehensively solve these challenges. Firstly, in order to improve the effectiveness of multi-target features, we adopt the encoder-decoder structure as the backbone segmentation framework and design a feature aggregation module (FAM) to realize the multi-level feature representation ($Stage1$). Secondly, because the segmentation results obtained from $Stage1$ are limited to the decoding of single scale feature maps, we design a multi-level attention mechanism (MLAM) to assign more attention to the multiple targets, so as to get multi-level attention maps. We fuse these attention maps and concatenate the output of $Stage1$ to carry out the second segmentation to get the final segmentation result ($Stage2$). The proposed method has better segmentation performance and balance on 2017 MM-WHS multi-modality whole heart images than the state-of-the-art methods, which demonstrates the feasibility of TSFM-Net for accurate segmentation of heart images.

1. Introduction

Cardiovascular diseases have increasingly become one of the main killers of human health, its morbidity and mortality are more than neoplastic diseases and leaped to the first place in the world [21]. It is an ischemic or haemorrhagic disease of the heart caused by hyperlipidemia, arteriosclerosis, hypertension, etc. Therefore, accurate analysis of the entire heart underlying structure is an important prerequisite for disease diagnosis, pathological analysis and surgical planning. In clinical practice, computed tomography (CT) and magnetic resonance imaging (MRI), as major non-invasive imaging techniques, can detect and capture abnormal changes in heart structures from multiple angles [14]. However, due to the individual differences and dynamic changes of heart structures, early cardiac image segmentation heavily relied on the manual work of experienced experts, which is time-consuming, error-prone, and subject to the subjective influence of operators [23]. Unfortunately, it is impossible to judge the abnormal heart structures only by manual work of cardiovascular CT and MRI images in clinical practice. These limitations have stimulated a large number of researches on full-automatic segmentation of whole heart images.

Deep convolutional neural networks (DCNNs) have achieved impressive effects in the field of medical image processing, especially in medical image segmentation [35, 43, 44]. For whole heart CT and MRI image segmentation, many automatic segmentation methods based on DCNNs have been proposed. For example, the two-stage 3D U-Net framework proposed by Wang et al. [31] can achieve multi-category image segmentation. Payer et al. [24] adopted the method of location first and then segmentation, and CNN is used for location and segmentation. Mortazi et al. [22] proposed an adaptive fusion strategy based on multi-plane CNNs to automatically fuse the parallel information of 3D scanning plane. Yang et al. [41] proposed hybrid loss guided deep supervised convolutional neural network, including weighted crossentropy (Wcross) loss and multi-class dice similarity coefficient (mDSC) loss. Although these methods have obtained general segmentation results, they did not take enough into account the inherent segmentation challenges of cardiac structures.

These challenges can be summarized in two ways. First, cardiovascular images contain multiple cardiac substructures, their pixel intensity has the situation of inconsistent within the classes and consistent between the classes, so it is difficult for simple DCNNs to capture discriminative multi-target features. Meanwhile, both target ambiguity and boundary ambiguity also make DCNNs easy to confuse the characteristic difference between target and background. Second, in the cardiovascular images, each slice contains different number of cardiac substructures, which leads to an extreme imbalance in the number of training samples for multi-targets. Such data imbalance makes it difficult to maintain a uniform level of segmentation performance for each cardiac substructure. Figure 1 shows some representative CT slices to elaborate on these challenges. The yellow circle and arrow in Figure 1(a) represent the area with fuzzy boundaries, Figure 1(b) points to the target area is vague.

*This work was supported in part by the Provincial Natural Science Research Program of Higher Education Institutions of Anhui province under Grant KJ2020A0035.

ORCID(s):

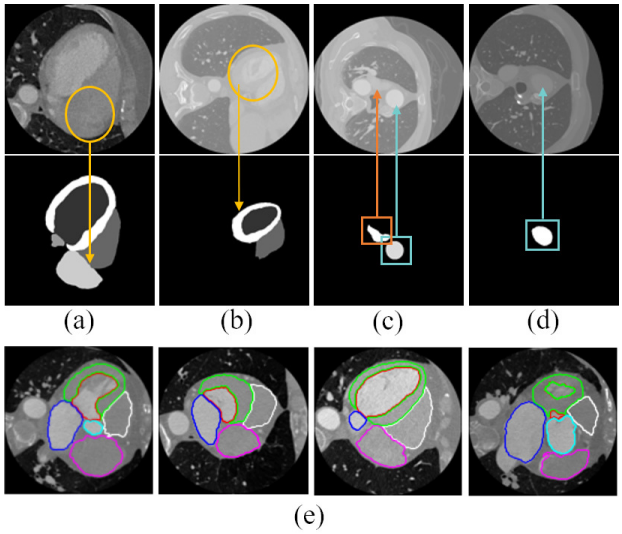


Figure 1: Examples of whole heart CT images. (a) The boundaries between the targets are blurred. (b) The targets are fuzzy. (c)-(d) The sample size is unbalanced. (e) Variability and diversity in shape and appearance of the targets.

Figure 1(c) and (d) represent the slices only have less targets. The four slices in Figure 1(e) show the contours of the multiple targets, which indicates the diversity of cardiac substructures in shape and appearance. For instance, the red contour area represents the left ventricle, but the left ventricle in the four slices varies in pixel intensity, shape and size. These phenomenons also exist for MRI images.

Among these challenges, the sample imbalance problem needs to be addressed emphatically. When processing multi-target images, DCNNs cannot adaptively select correlation features in response to the changes of multiple targets. In most cases, DCNNs treat all the objects in an image equally and capture their spatial features through the convolutional layers. As a result, more features are extracted for the objects with larger size, and vice versa. At this time, if the sample number of big target is more and the sample number of small target is few, the training will be seriously unbalanced. To solve this problem, many scholars have proposed some feature extraction schemes based on attention models [32, 33, 15, 8]. For instance, Wang et al. [32] proposed a CNN using residual attention mechanism to generate attention perception features by stacking attention modules. Wang et al. [33] proposed a deep spatial attention module to extract 2D and 3D image features. There are also some attention models aimed at scene segmentation, such as the feature pyramid attention module proposed by Li et al. [15], the position attention module and channel attention module proposed by Fu et al. [8]. The application of these attention models enables DCNNs to allocate more attention to important areas of an image. Inspired by this, we also try to design a feature aggregation module and a multi-level attention mechanism to make DCNNs pay more attention to each target, so as to improve the sample balance of multiple targets.

In this paper, we propose a two-stage segmentation network with feature aggregation and multi-level attention mechanism (TSFM-Net) to segment cardiac substructures in whole heart CT and MRI images. In the first stage, we design a feature aggregation module (FAM), which generates high-level feature maps at different levels by multi-scale maxpooling operation at each scale feature map. These high-level feature maps are the effective summaries of the local key information in the low-level feature maps, so the global key information feature maps are formed when the high-level feature maps of same scale are fused. Then, the semantic feature maps are generated by continuous upsampling layers. Such encoder-decoder structure with FAM produces the segmentation masks of the first stage ($Stage1$). In $Stage1$, FAM only provides multi-level features for multi-target segmentation at the feature level. However, the imbalance among the multiple targets still needs to be improved. Thereby, in the second stage, we propose a multi-level attention mechanism (MLAM) to assign attention to target areas, so as to give more attention to each target channel. And then immediately, the attention maps are sent into attention fusion module (AFM) for the second segmentation ($Stage2$). In this way, $Stage1$ focuses on the whole target region, while $Stage2$ focuses on each target channel based on the multi-level attention maps. This complementary approach reduces the interference and imbalance among the multiple targets.

To sum up, the main contributions of this paper are as follows:

(1) We propose a two-stage segmentation network with feature aggregation and multi-level attention mechanism (TSFM-Net) for multi-target segmentation task, which is used to solve the unbalance and disturbance of a single end-to-end segmentation network for multi-target segmentation.

(2) We design a feature aggregation module (FAM) for extracting multi-level features. Through the multi-scale maxpooling layers, the advanced feature maps containing comprehensive information of different levels can be obtained for the guidance of segmentation.

(3) We creatively propose a multi-level attention mechanism (MLAM) to focus on the target areas in the intermediate feature maps. Bilateral attention modules (BAM) improve the salience of the target by using the attention mechanism on the compressed channel feature vectors.

The rest of this paper is organized as the following: In Section 2, the related works will be introduced. In Section 3, we will give a detailed description of the proposed segmentation framework. In Section 4, the datasets, experimental setups and evaluation indexes will be introduced. Particularly, the comparative experimental results and ablation studies are presented and analyzed emphatically in this section. Finally, the conclusions of this paper are given in Section 5.

2. Related work

2.1. Deep learning for cardiac structure segmentation

In recent years, due to the end-to-end training advantages and superior feature discrimination ability, deep learning algorithms gradually replace traditional machine learning methods [46, 13, 9], especially in cardiac structure segmentation [38, 39, 37]. Both Luo et al. [17] and Payer et al. [24] adopted CNN to achieve location and segmentation. Full convolutional neural network (FCN) and its variants have also been actively used in the segmentation of cardiac images. For example, Qin et al. [26] designed a network of simultaneous cardiac motion estimation and segmentation based on FCN. The recursive full convolutional network (R-FCN) proposed by Poudel et al. Poudel et al. [25] and Xue et al. [40] could capture the spatial dependence between slices through the memory unit while learning the image representation. Nicoló et al. [29] proposed the application of generative adversarial network (GAN) in the full-resolution image and region of interest (ROI), which better guides the learning of FCN through the mutual learning of the two types of resolution images. The above works only achieve limited segmentation performance, because they simply transplant the classical deep learning algorithms from computer vision field to cardiac image segmentation task, which greatly ignores the specificity and challenges of medical images.

To better match the deep learning algorithms to cardiac images, many scholars have developed segmentation methods aiming at the motion and morphological characteristics of cardiac structures. Shi et al. [30] proposed the probabilistic deep voxel dilated residual network and Khened et al. [19] proposed the multi-scale residual DenseNets. They mainly use the residual networks to deal with the morphological changes and differences of multi-target images to avoid the loss of features. Du et al. [5] proposed an improved inception module. Through the multi-scale convolutional layers and 1×1 convolutional layers, the extracted image features are more hierarchical and have stronger expression ability. Densely expanded convolution and feature pyramid network can also extract multi-level features well. Zheng et al. [45] proposed heterogeneous feature aggregation network (HFA-Net), which can make full use of complementary information and clustered heterogeneous characteristics in 3D cardiac images to improve segmentation performance. Fei et al. [7] used the local attention networks with pyramid structure for feature maps of different levels and carried out consistency learning to pay attention to the cardiac structures. All of these methods show superiority in cardiac structure segmentation.

2.2. Attention mechanism

Attention mechanism allows the network to learn to focus on important information and ignore irrelevant information, its application to medical image has achieved remarkable results. Yi et al. [42] proposed to add an attention model to the feature map at each scale in the decoding block to realize the localization and segmentation of cell images. Wang et al. [34] proposed three attention models for the classification of 14 thoracic diseases, including channel-wise attention, element-wise attention and scale-wise attention. Chen et al. [4] proposed the attention mechanism to generate new labels based on background and target respectively for soft segmentation, and these labels are used to assist the training of segmentation network. Fei et al. [7] proposed a local attention network with pyramid structure, which gives local attention to full-resolution feature maps of different levels and integrates them into global attention maps. Xing et al. [36] proposed the attention-guided deformation module, which minimizes the interference caused by deformation in the process of enlarging the lesion areas. These methods all used attention mechanism to highlight the target areas for different points.

2.3. Multi-level feature extraction

Many researchers have done constructive works in image feature extraction. The single convolutional layer and pooling layer usually cannot provide sufficient and valuable feature information for multi-target image segmentation, which leads to many effective methods of multi-level feature extraction. For example, Mathai et al. [20] and Rad et al. [27] directly used multi-scale convolutional layers. Gao et al. [10] and Ma et al. [18] both extracted multi-level features by overlaying a dilated spatial pyramid pooling module in the coding block. Fang et al. [6] respectively sent a single image and image sequence into the parallel convolutional network, then extracted the spatial features and temporal correlations of the images and fused them. Boot et al. [2] utilized the structure of three-step convolution and downsampling/upsampling combined with skip connections to make the features more refined and have stronger expression ability. Fei et al. [7] extracted multi-level feature information by downsampling at different scales for feature maps of different levels. However, although the success of these methods, they did not take into account the different feature levels of multi-target images, and the relationship between local features and global features at the same time.

3. Methodology

3.1. Task description

The goal of our work is to automatically segment 7 cardiac substructures in whole heart CT and MRI images by an end-to-end deep learning network without manual intervention. Specifically, in our method, the input is a single CT or MRI image with the size of 128×128, the output is the segmentation mask with the same size. As shown in Figure 2, the segmentation process is mainly achieved by two stages. $Stage1$ not only output the segmentation masks, but also generates the bilateral feature maps. In $Stage2$, these intermediate feature maps are transformed into multi-level attention maps. The final segmentation results are produced through the fusion and connection with the segmentation maps of $Stage1$. All the segmentation processes are based on the classification at the pixel level. For pixel $x^{(i)}$, its

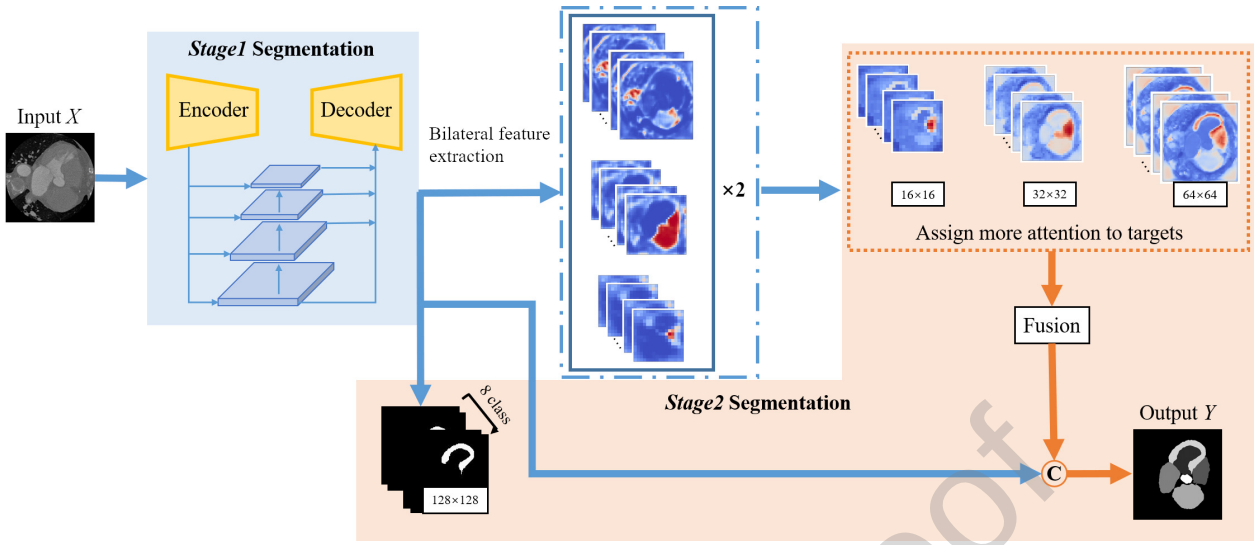


Figure 2: Task description diagram.

classification result can be defined as

$$P\left(y^{(i)}=j|\,x^{(i)};\theta\right)=\frac{e^{\theta_j^T x^{(i)}}}{\sum_{l=1}^{k} e^{\theta_l^T x^{(i)}}} \tag{1}$$

$$P\left(x^{(i)}\right)=\max imum\left(P\left(y^{i}=j\right),j\in[1,k]\right) \tag{2}$$

that is, $P\left(y^{(i)}=j|\,x^{(i)};\theta\right)$ is the probability of using softmax to classify $x^{(i)}$ as category j. k is the number of possible categories. The category with the highest probability $P\left(x^{(i)}\right)$ should be taken as the final category of pixel $x^{(i)}$.

3.2. Method overview

The proposed two-stage segmentation network (TSFM-Net) is shown in Figure 3. Its input is a CT or MRI image with resolution of $H\times W$ (H=128, W=128). In $Stage1$, original images are sent into an encoder with feature aggregation module (FAM), which contains four feature extraction blocks to extract multi-scale feature information. The multi-scale maxpooling layers are used to filter the feature maps of each scale, so that the generated high-level feature maps are the local generalization at different levels of the low-level feature maps. Then the advanced feature maps of same scale are aggregated to form the global feature expression of the original image. In the process of downsampling and upsampling, three kinds of bilateral feature maps (16×16, 32×32, 64×64) are produced, which provides guidance for $Stage1$ segmentation. In $Stage2$, bilateral feature maps are sent into a multi-level attention mechanism (MLAM) to create attention maps. The classification accuracy can be improved by increasing the focus on the target areas. Three bilateral feature maps have ability to provide different levels of attention to the target while preserving enough information of the encoding module. By this way, the multi-level attention maps could be more similar to the real masks. Subsequently, we scale up these attention maps through step-by-step upsampling operation. At the same time, the low-resolution attention maps are transferred to the high-resolution attention maps by attention fusion module (AFM), so as to achieve the purpose of integrating scale differences. These attention maps and the segmentation maps of $Stage1$ construct the segmentation masks of $Stage2$.

The output feature maps of $Stage1$ and $Stage2$ all have 8 channels which consist of a background channel x_0 and 7 channels x_1, x_2, x_3, x_4, x_5, x_6, x_7 corresponding to the 7 targets. Such results are realized by Softmax classification function, that is, Softmax calculates the feature vector at each spatial position of the input feature maps and obtains 8 probability values, which represent the probability that the pixel at this position belongs to these 8 categories. We further employ the categorical cross-entropy loss to supervise the training of $Stage1$ segmentation network and $Stage2$ segmentation network. The loss function is described as

$$L_i=-\sum_{t=1}^{8} y_t \log\left(p_{t_i}\right), i=Stage1, Stage2 \tag{3}$$

where t represents the target class, and there are 8 classes of targets together with the background. y_t is the label. If y_t=1, it belongs to class t, otherwise it is 0. p_t is the output of the neural network, that is, the probability of category t. This output value is calculated by Softmax. The total loss of the proposed segmentation framework is $L_{Total}=L_{Stage1}+L_{Stage2}$.

3.3. Stage1 segmentation network

3.3.1. Feature aggregation module

We propose a feature aggregation module (FAM) as shown in Figure 4, whose function is to aggregate high-level feature information of the same scale. In the four encoding blocks, we use two 3×3 convolutional layers to extract more

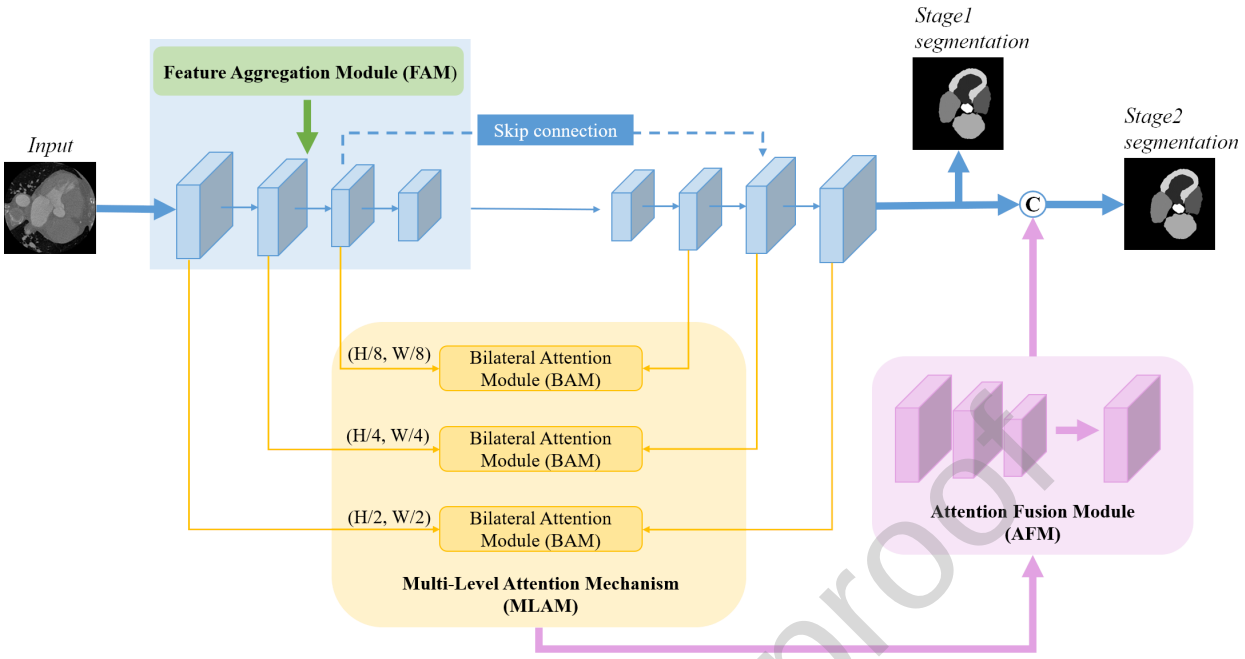


Figure 3: Modular description of the two-stage segmentation network (TSFM-Net).

expressive feature information. Starting from the first convolutional block, we use the maxpooling operation at different scales (2×2, 4×4, 8×8, 16×16) for each feature map, which results in some abstract feature maps with lower resolution. Set the output of the convolutional layer is a^{l-1}, the output of the pooling layer is a^{l}, the maxpooling formula is

$$a_{ij}^{l} = \max\left(a_{mn}^{l-1}\right), i \le m, n \le i + k, k = 1, 3, 7, 15 \quad (4)$$

where m, n represent the region covered by the pooling kernel.

High-level feature maps with different resolutions are the abstract representations of the original image. For the same feature map, the maxpooling operations with different kernel sizes result in different levels of advanced feature maps. The larger the pooling kernel is, the more global information is attached to the extracted feature map. While the smaller the pooling kernel is, the extracted feature map focuses more on local information. Then the feature maps of the same resolution are concatenated as input for the next encoding block. The fusion formulas for each block are: $F^1 = a^{l1}$, $F^2 = concate(a^{l1}, a^{l2})$, $F^3 = concate(a^{l1}, a^{l2}, a^{l3})$, $F^4 = concate(a^{l1}, a^{l2}, a^{l3}, a^{l4})$. In this way, the multi-level features can be retained without loss of information, so as to form the global feature expression of the original image in the continuous fusion process of local features.

3.3.2. Stage1 network structure

*Stage*1 segmentation network is mainly a symmetrical U-shaped structure as shown in Figure 3, with encoder and decoder as the main body. Table 1 shows the structure and parameter settings of *Stage*1 segmentation network. It can be seen from the table that segmentation is achieved by encoding and then decoding the input image. The encoding part is FAM, which has four encoding blocks with the same structure, each block is composed of two convolutional layers, a pooling layer and a dropout layer. The decoding part is also composed of four identical upsampling blocks, each of which consists of an upsampling layer, two convolutional layers and a dropout layer. To prevent network from tending to overfitting state under the limitation of training sets, we set the value of dropout layer to 0.35 to reduce the parameters of the previous layer. Softmax function is set in the last layer of the decoding section to calculate the probability that each pixel belongs to eight categories and then the category with the highest probability is selected as the final classification of the pixel. The essence of multi-target image segmentation is to classify each pixel and finally form a target classification map with 8 channels.

3.4. Stage2 segmentation network

3.4.1. Multi-level attention mechanism

Bilateral feature maps containing useful guiding information extracted from the encoder and decoder of *Stage*1 segmentation network. They provide guidance for pixel classification and semantic segmentation at different levels. Specifically, the encoding features preserve detail information, while the decoding features preserve more semantic information. To a large extent, these feature maps can be considered as a high generalization of the objects, but there is still a lot of interference information.

For better and more attention to the target areas, a multi-level attention mechanism (MLAM) is constructed for generating attention maps, which is composed of three bilateral attention modules (BAM) corresponding to the bilateral feature maps of three scales. As shown in Figure

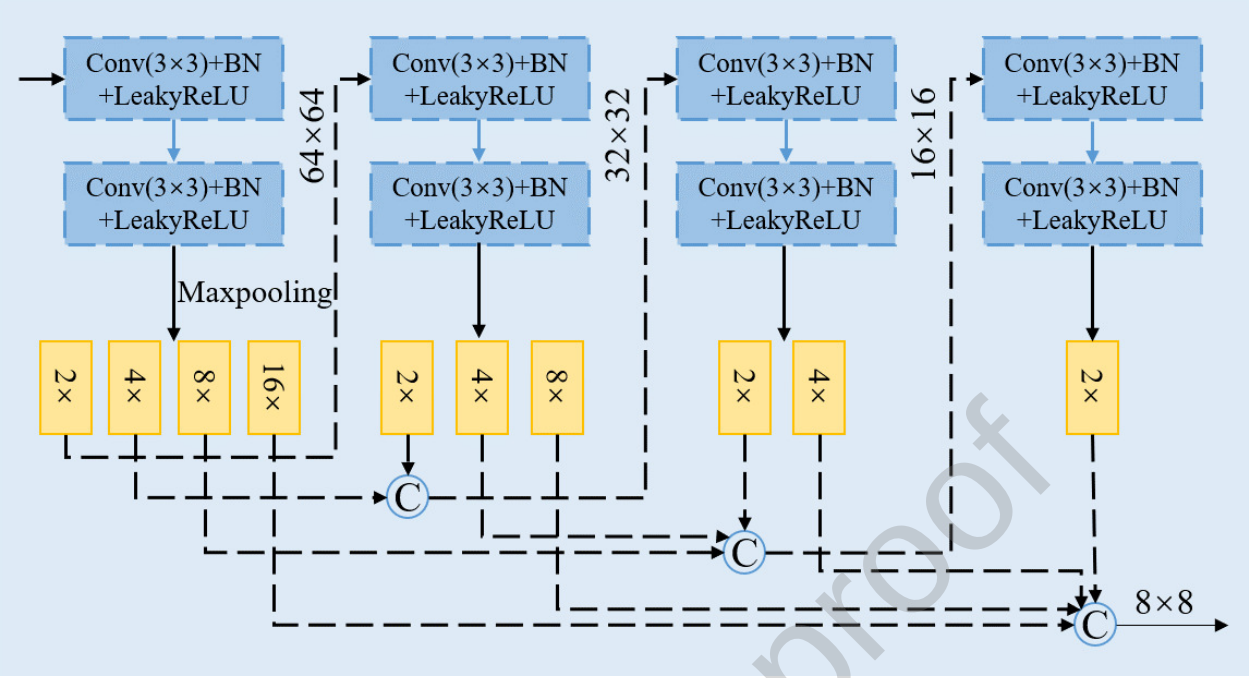


Figure 4: Illustration of feature aggregation module (FAM).

Table 1
Structure and parameter settings of *Stage*1 segmentation network.

Encoding									Decoding			
	Input		128×128			Output						Output
	Convolution		3×3			128×128				Upsampling	2×2	16×16
Blcok1			3×3							Convolution	3×3	
	Maxpooling	2×2	4×4	8×8	16×16	64×64	32×32	16×16	8×8		3×3	
	Dropout		0.35							Dropout	0.35	
	Convolution		3×3			64×64				Upsampling	2×2	32×32
Blcok2			3×3							Convolution	3×3	
	Maxpooling	2×2	4×4	8×8		32×32	16×16	8×8			3×3	
	Dropout		0.35							Dropout	0.35	
	Convolution		3×3			32×32				Upsampling	2×2	64×64
Blcok3			3×3							Convolution	3×3	
	Maxpooling	2×2	4×4			16×16	8×8				3×3	
	Dropout		0.35							Dropout	0.35	
	Convolution		3×3			16×16				Upsampling	2×2	128×128
Blcok4			3×3							Convolution	3×3	
	Maxpooling		2×2			8×8					3×3	
	Dropout		0.35							Softmax		128×128×8

5, we first use the globalaveragepooling layer and the globalmaxpooling layer to compress the feature maps. Both of these two pooling ways have the ability to represent a feature map in the form of a pixel value, but the expression degree of feature map is different. Globalmaxpooling takes the largest pixel value in the feature map as the global representation, while globalaveragepooling takes the average value of pixel values as the global representation. Among them, the feature vector obtained from the globalmaxpooling layer has a strong correlation with the targets. We perform attention calculation on the bilateral feature vectors to highlight the target intensity. Furthermore, deeper feature learning and expression are also achieved through the fully connection layer and ReLU. For the feature vectors from globalmaxpooling layer, the interaction between them is to deepen the attention to the efficient channel features. As for the feature vectors from globalaveragepooling layer, attention calculation is carried out in an independent form, which

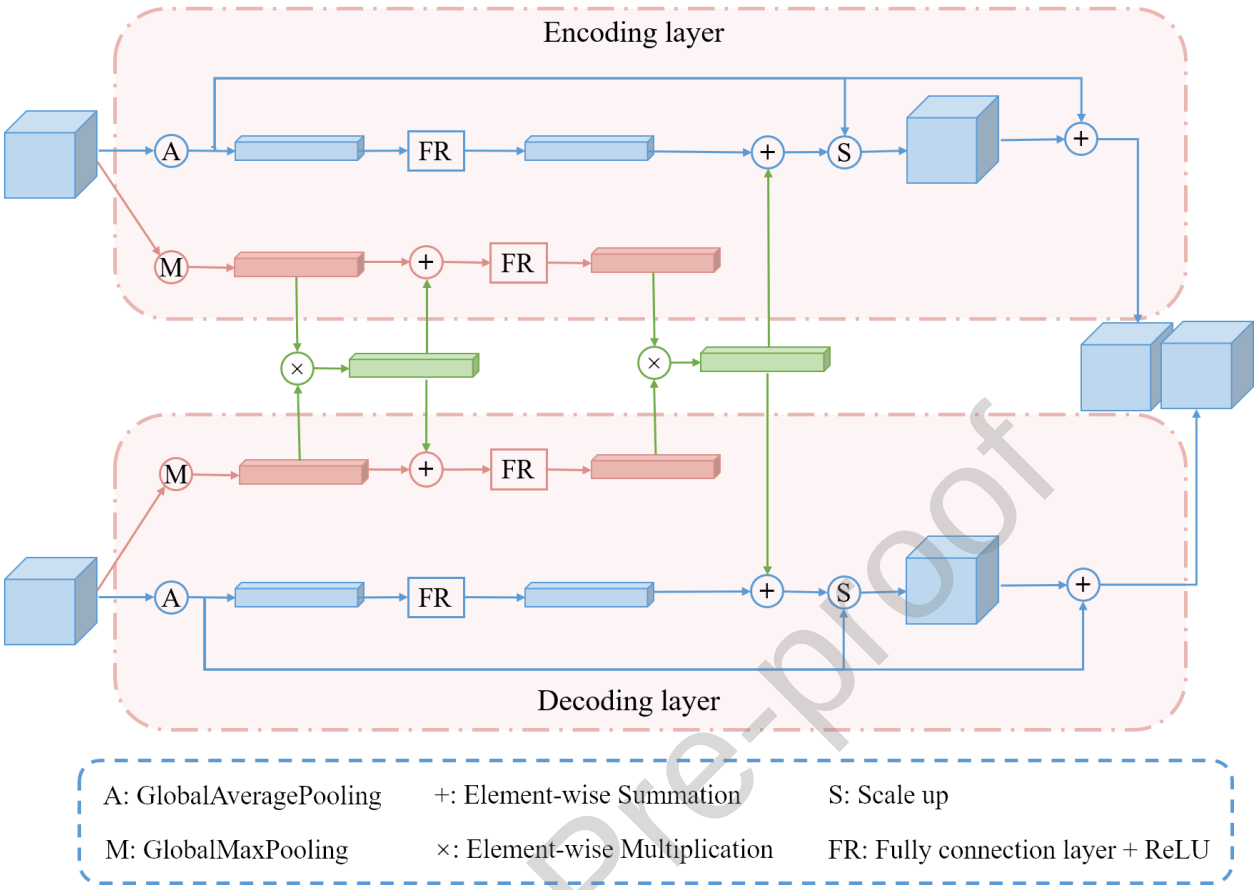


Figure 5: Illustration of bilateral attention module (BAM).

could ensure that the specificity of the information contained in them remains unchanged. The concatenation of the two attention maps has the ability to make the target areas more obvious and suppress the background areas.

3.4.2. Attention fusion

We propose a two-stage segmentation strategy to make up for the limitation of *Stage*1 segmentation network in terms of scale variation. As shown in Figure 6, we show the fusion process of the multi-level attention maps. Because *Stage*1 segmentation masks are obtained from step-by-step upsampling operation of single scale feature maps, the key features are easy to be diffused in the process of scaling up. Hence, for the purpose of retaining the important information of each resolution, we send the bilateral feature maps into MLAM to make the targets more obvious in the feature maps and form the multi-level attention maps. What cannot be ignored is that, these attention maps contain enough structural information but also lacks sufficient context information. To solve this problem, we design an upsampling branch and add short connections to transfer low-resolution features to high-resolution feature maps. Namely, attention fusion module (AFM). These three resolution attention maps complement each other at the feature level and form sufficient context connection.

4. Experiments

4.1. Dataset

The proposed method (TSFM-Net) is validated on whole heart multi-modality images from MM-WHS 2017 Challenge dataset. The goal of the challenge is to segment 7 cardiac substructures from CT and MRI images, including the left ventricle blood cavity (LV), the right ventricle blood cavity (RV), the left atrium blood cavity (LA), the right atrium blood cavity (RA), the myocardium of the left ventricle (LV_Myo), the ascending aorta (AA) and the pulmonary artery (PA). The cardiac CT data are obtained using two advanced 64-slice CT scanners (Philips Medical Systems, Netherlands) using a standard coronary CT angiography protocol at two sites in Shanghai, China. All image data from the upper abdomen to the aortic arch covers the whole heart. The in-plane resolution of the axial section is 0.78×0.78 mm and the average section thickness is 1.60 mm. The cardiac MRI data are obtained from two hospitals in London, England. One set of data is taken from a 1.5-T Philips scanner, the other from a 1.5-T Siemens scanner. Data are collected at a resolution of approximately (1.6∼2)×(1.6∼2)×(2∼3.2) mm and reconstructed to half of its acquisition resolution, i.e., about (0.8∼1)×(0.8∼1)×(1∼1.6) mm.

The CT and MRI volume images are sliced, removed all-black label images and original images. The number

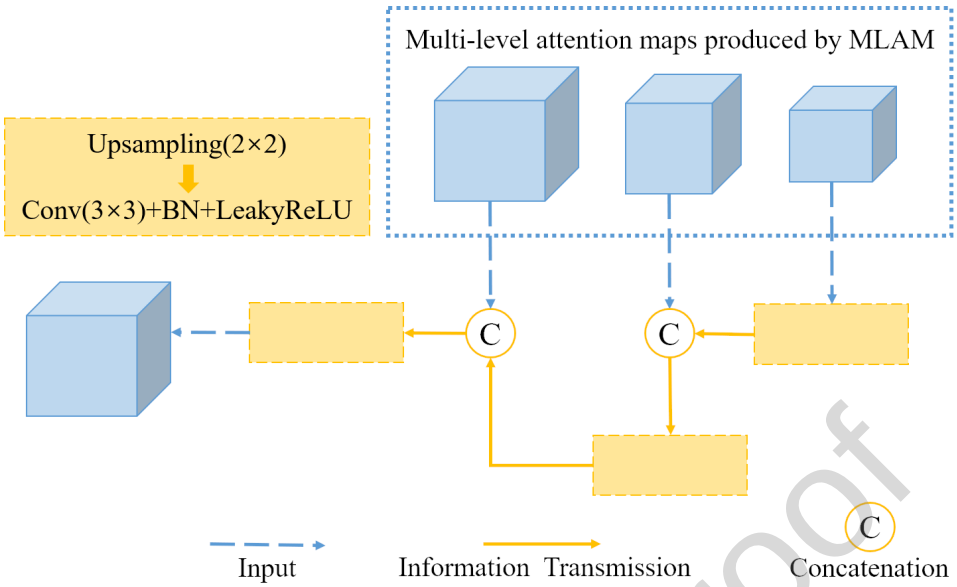


Figure 6: Illustration of attention fusion module (AFM).

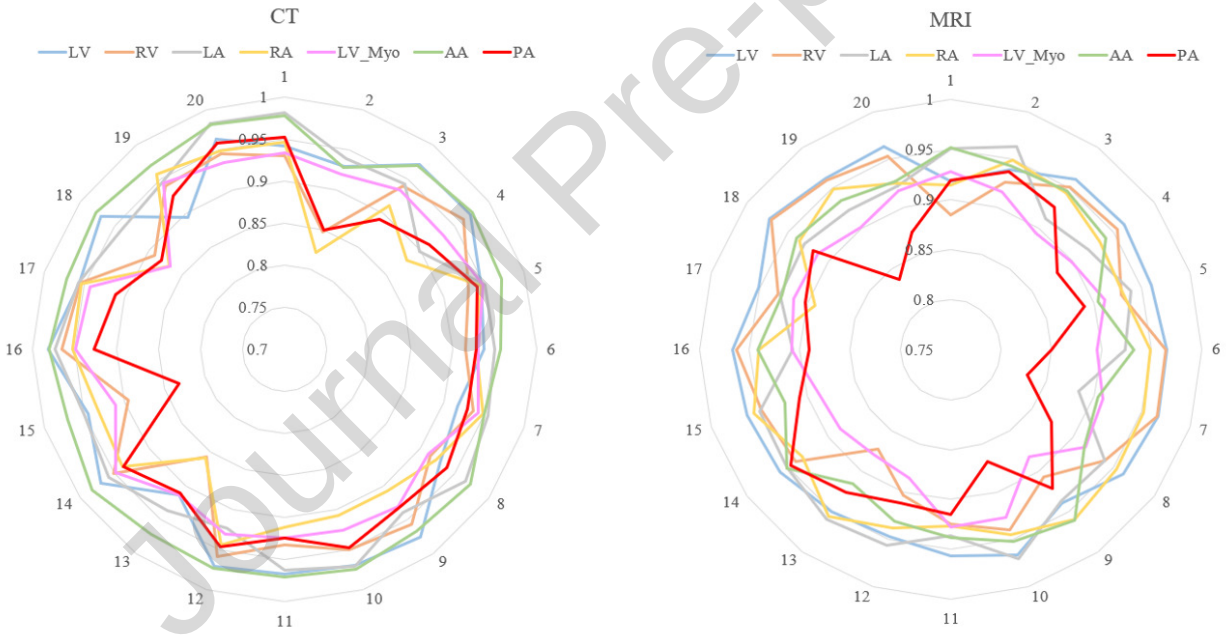


Figure 7: Segmentation results (Dice) of 20 subjects.

of slices retained varies from patient to patient, ranging from 161 to 357 of CT and 60 to 175 of MRI. The high degree of inconsistency in the number of patients'slices brings great difficulty to our experiment. To reduce the computational cost and time, we resize all the slices to resolution of 128×128. Meanwhile, the pixel values of each slice are normalized to [0, 1]. For the sake of the stability and robustness of segmentation model, the training samples are augmented through rotation.

4.2. Implementation details

Our deep learning segmentation model is implemented using Keras library on TensorFlow platform. The entire experiments are carried out with Intel Core i7-7820×3.60GHz processor and an Nvidia GeForce GTX 1080 TI with 32GB of RAM. The training images are processed into normalized images with resolution of 128×128 by MatLAB. We set the learning rate as 0.001 (Adaptive optimizer) and mini-batch as 16. Categorical crossentropy is used as the classification loss function of supervised learning. *Stage*1 segmentation network is trained firstly, then we apply the segmentation

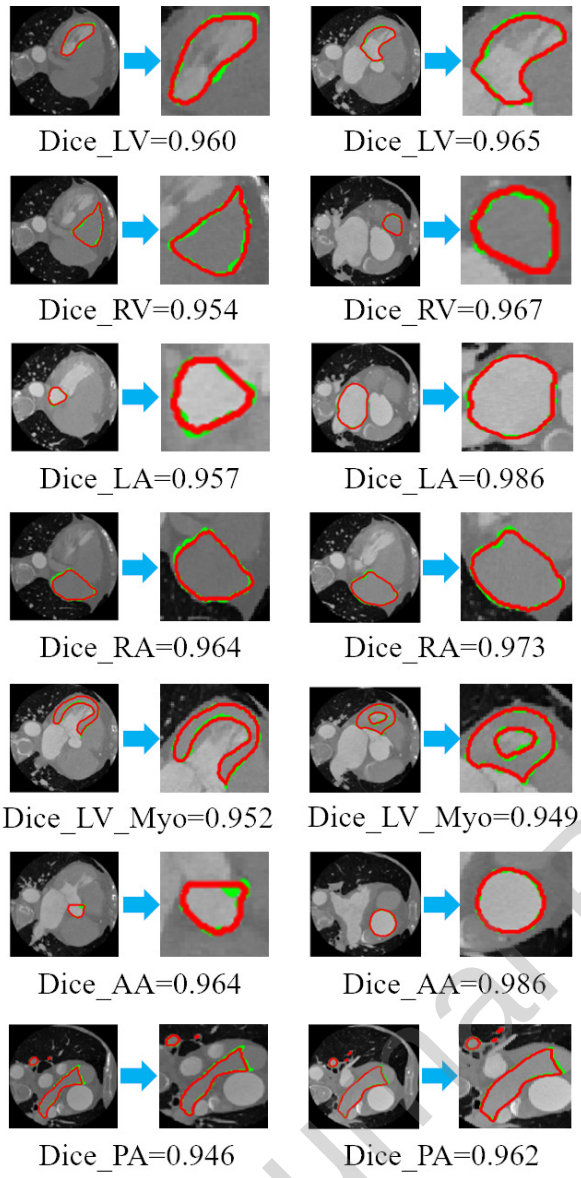


Figure 8: Visualization of segmentation results.

maps and the bilateral feature maps to *Stage*2 segmentation network. A total of 100 iterations of training are conducted. Since this public challenge only provide training sets and labels of 20 subjects, we use 5-fold cross-validation method to verify the effectiveness of our segmentation framework. Each training has 16 training sets (training sets accounts for 80%, validation sets accounts for 20%) and 4 test sets, the whole experiment is run 5 times. The average value of all the test sets is taken as the final segmentation result.

4.3. Evaluation metrics

To make a fair comparison with the state-of-the-art segmentation methods, Dice correlation (Dice) is considered to evaluate the segmentation performance of 7 substructures:

$$Dice(P,G) = 2 \times \frac{P \cap G}{P + G} \tag{5}$$

where P represents the predicted segmentation results, G represents the ground-truth segmentation results. For ablation study, we also use ROC (Receiver Operating Characteristic) curve and AUC (Area Under the Curve) value to evaluate the rationality and validity of the proposed segmentation network. The calculation formula of AUC is

$$AUC = \frac{\sum_{pro_i \epsilon positiveclass} rank_{pro_i} - \frac{P \times (P+1)}{2}}{P \times N} \tag{6}$$

where $rank_{pro_i}$ represents the serial number of the ith sample, $\sum_{pro_i \epsilon positiveclass} rank_{pro_i}$ represents the summation of the numbers of the positive samples. P and N represent the number of positive samples and the number of negative samples.

4.4. Experimental results and analysis

4.4.1. Segmentation performance evaluation

The segmentation results of 20 subjects are presented in Figure 7. For CT images, AA has the best segmentation result and the average Dice value can reach 0.973. Moreover, the segmentation performance of AA can be maintained at a stable level for each subject. Of course, for the other 6 targets, there are some poor segmentation of individual subjects, which affect the overall segmentation level. For MRI images, LV has the highest average Dice value can reach 0.959 and the best stability. As can be seen from Figure 8, the predicted contours of images with high Dice values highly coincide with the real contours. Through visualization, we find that when the target is larger, the segmentation performance is better, while when the target is smaller, the segmentation performance is slightly worse. Hence, in the segmentation of multi-target images, the size of the target has a serious impact on the segmentation performance.

4.4.2. Compared with the state-of-the-art methods

We directly compare and analyze the proposed method with the state-of-the-art methods. These methods have worked well in this challenging dataset. Table 2 and 3 show the segmentation results of CT images and MRI images respectively. Our method achieves the best results on 7 targets on both modality images. For CT images, Gu et al. [11] obtained the best quantitative results on the segmentation of LV and AA, but the mean Dice value of TSFM-Net can reach 0.940. For MRI images, although Wang et al. [31] achieved better segmentation results on LV_Myo and PA than ours, our experimental results are absolutely competitive on the whole and the mean Dice value can reach 0.934. It can be seen from Table 2 and 3 that these methods have a certain degree of imbalance in the segmentation of the 7 objectives. For example, Yang et al. [41] achieved a Dice value of 0.941 for AA in CT images, but the results on other targets are very poor. Mortazi et al. [22] can obtain the Dice value of 0.932 for LV in MRI images, but the results of other targets also seriously lower the average value. Here, we also show the imbalance among the segmentation results of the 7 targets in the form of standard deviation in Table 4 and 5. The standard deviation value of our method can reach 0.019 on CT and

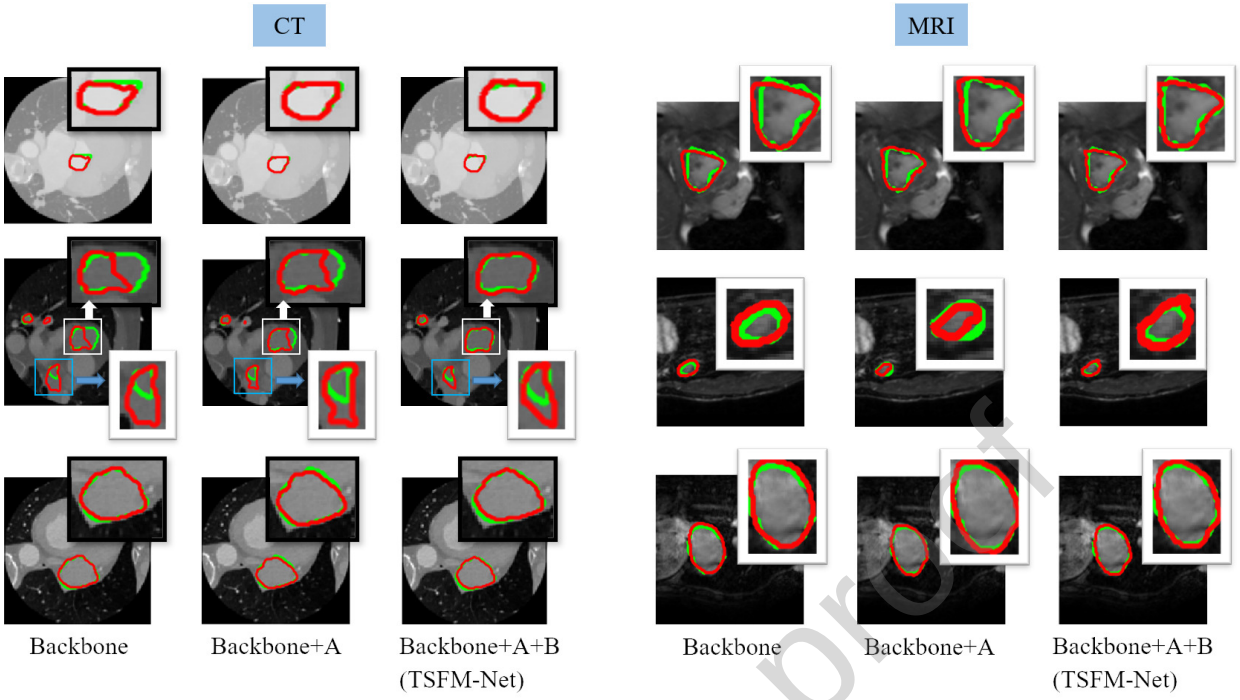


Figure 9: Segmentation result visualization of TSFM-Net and the ablation networks.

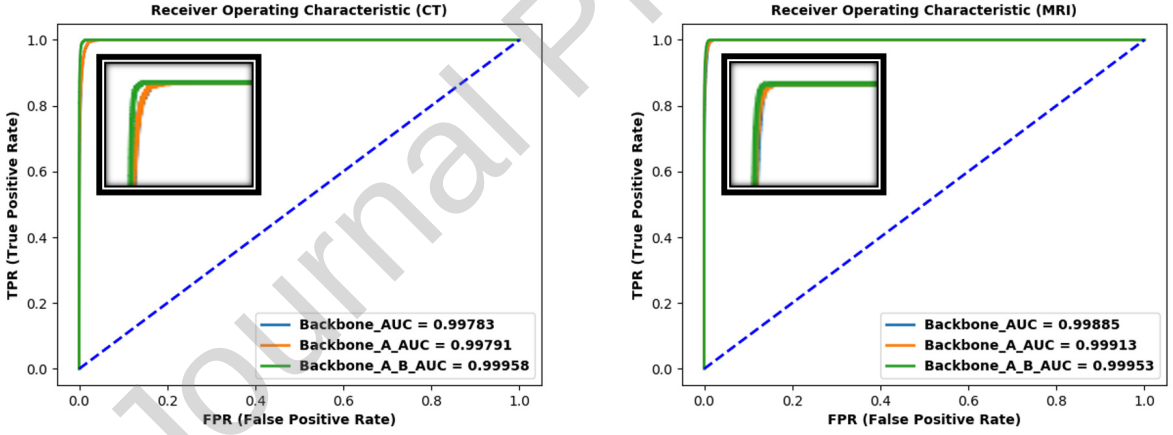


Figure 10: AUC and ROC curves of TSFM-Net and the ablation networks.

MRI images. It is superior to the state-of-the-art methods.

As for the huge differences in the segmentation results of the 7 targets, we believe that this has a lot to do with the processing of the dataset. For 3D CT and MRI volume images, slices from three angles can be selected for 2D experiments. The slices from different angles as the input images inevitably lead to certain differences in the number of samples. In addition, the experiments are directly carried out on 3D images are also different from 2D images. In these comparative experiments, HFA-Net (DVN, S-DVN) [45] directly used 3D volume images as the input of the segmentation network. It is worth noting that, although the data preprocessing methods are different, our method still have the ability to achieve a relatively stable segmentation performance for all targets without bad segmentation results, which is the advantage of TSFM-Net using FAM, MLAM and the two-stage segmentation strategy.

4.4.3. Compared with the benchmark methods

In our method, *Stage*1 segmentation framework is a U-shaped structure including feature aggregation module (FAM) and skip connections. Therefore, we need to reproduce the benchmark segmentation networks such as FCN [16], U-Net [28], SegNet [1], ResNet50 [12]. Table 6 and 7 show the quantitative segmentation results of these benchmark networks. These benchmark methods are all segmentation based on pixel-level classification, which is very similar

Table 2
Comparison of TSFM-Net and the state-of-the-art methods on CT images (Dice).

Method	LV	RV	LA	RA	LV_Myo	AA	PA	Mean
Wang et al. [31]	0.837	0.860	0.861	0.862	0.859	0.918	0.885	0.869
Payer et al. [24]	0.918	0.909	0.929	0.888	0.881	0.933	0.840	0.900
Mortazi et al. [22]	0.930	0.888	0.925	0.877	0.898	0.909	0.851	0.897
Yang et al. [41]	0.878	0.778	0.845	0.815	0.819	0.941	0.826	0.843
DVN [45]	0.942	0.891	0.933	0.879	0.908	0.959	0.824	0.905
S-DVN [45]	0.929	0.890	0.914	0.899	0.895	0.956	0.828	0.902
HFA-Net [45]	0.946	0.893	0.925	0.897	0.910	0.964	0.830	0.909
Chen et al. [3]	0.919	——	0.911	——	0.877	0.927	——	0.909
Gu et al. [11]	**0.951**	0.902	0.938	0.911	0.922	**0.974**	0.837	0.919
TSFM-Net	0.950	**0.936**	**0.951**	**0.922**	**0.926**	0.973	**0.922**	**0.940**

Table 3
Comparison of TSFM-Net and the state-of-the-art methods on MRI images (Dice).

Method	LV	RV	LA	RA	LV_Myo	AA	PA	Mean
Payer et al. [24]	0.916	0.868	0.855	0.881	0.778	0.838	0.731	0.838
Mortazi et al. [22]	0.932	0.884	0.887	0.874	0.825	0.772	0.784	0.851
Yang et al. [41]	0.858	0.819	0.787	0.820	0.742	0.726	0.698	0.779
Shi et al. [30]	0.914	0.880	0.856	0.873	0.811	0.857	0.794	0.855
Galisot et al. [9]	0.897	0.819	0.765	0.808	0.763	0.708	0.685	0.778
Heinrich et al. [13]	0.918	0.871	0.886	0.873	0.781	0.878	0.804	0.859
Chen et al. [3]	0.924	——	0.805	——	0.788	0.828	——	0.838
Wang et al. [31]	0.881	0.913	0.934	0.922	**0.938**	0.876	**0.921**	0.912
TSFM-Net	**0.959**	**0.943**	**0.940**	**0.945**	0.910	**0.935**	0.905	**0.934**

Table 4
Comparison of segmentation performance balance on CT images (Standard Deviation).

Method	Standard Deviation
Wang et al. [31]	0.026
Payer et al. [24]	0.033
Mortazi et al. [22]	0.028
Yang et al. [41]	0.053
DVN [45]	0.046
S-DVN [45]	0.040
HFA-Net [45]	0.043
Chen et al. [3]	0.022
Gu et al. [11]	0.044
TSFM-Net	**0.019**

Table 5
Comparison of segmentation performance balance on MRI images (Standard Deviation).

Method	Standard Deviation
Payer et al. [24]	0.063
Mortazi et al. [22]	0.059
Yang et al. [41]	0.058
Shi et al. [30]	0.041
Galisot et al. [9]	0.072
Heinrich et al. [13]	0.048
Chen et al. [3]	0.061
Wang et al. [31]	0.024
TSFM-Net	**0.019**

to our method in theory. As can be seen from the quantitative results in Table 6, the multi-target heart images cannot be accurately segmented only by the encoding-decoding structure.

4.4.4. Ablation Study

The proposed two-stage segmentation network is constructed by integrating multiple modules on the basis of encoder-decoder framework, that is backbone segmentation network. Hence, it is necessary to verify the effectiveness of each module on the segmentation performance through

Table 6
Comparison of TSFM-Net and the benchmark methods on CT images (Dice±Standard Deviation).

Method	LV	RV	LA	RA	LV_Myo	AA	PA	Mean
FCN	0.867 ±0.113	0.794 ±0.127	0.881 ±0.068	0.757 ±0.210	0.802 ±0.083	0.857 ±0.134	0.694 ±0.185	0.807 ±0.132
U-Net	0.900 ±0.025	0.854 ±0.053	0.929 ±0.032	0.862 ±0.032	0.818 ±0.051	0.907 ±0.085	0.726 ±0.174	0.857 ±0.065
SegNet	0.874 ±0.068	0.730 ±0.253	0.919 ±0.043	0.846 ±0.034	0.806 ±0.065	0.850 ±0.148	0.846 ±0.034	0.839 ±0.092
ResNet50	0.819 ±0.055	0.830 ±0.082	0.787 ±0.171	0.787 ±0.089	0.757 ±0.117	0.874 ±0.079	0.745 ±0.146	0.800 ±0.106
TSMLCA-Net	**0.950 ±0.025**	**0.936 ±0.026**	**0.951 ±0.022**	**0.922 ±0.034**	**0.926 ±0.014**	**0.973 ±0.012**	**0.922 ±0.034**	**0.940 ±0.019**

Table 7
Comparison of TSFM-Net and the benchmark methods on MRI images (Dice±Standard Deviation).

Method	LV	RV	LA	RA	LV_Myo	AA	PA	Mean
FCN	0.832 ±0.114	0.624 ±0.282	0.652 ±0.079	0.669 ±0.159	0.659 ±0.168	0.601 ±0.080	0.633 ±0.120	0.667 ±0.143
U-Net	0.898 ±0.042	0.773 ±0.117	0.808 ±0.033	0.785 ±0.116	0.758 ±0.032	0.717 ±0.050	0.739 ±0.091	0.783 ±0.069
SegNet	0.894 ±0.035	0.792 ±0.117	0.828 ±0.039	0.830 ±0.070	0.758 ±0.042	0.744 ±0.075	0.761 ±0.076	0.801 ±0.065
ResNet50	0.847 ±0.089	0.673 ±0.204	0.634 ±0.045	0.721 ±0.062	0.744 ±0.071	0.647 ±0.063	0.620 ±0.116	0.698 ±0.093
TSMLCA-Net	**0.959 ±0.012**	**0.943 ±0.023**	**0.940 ±0.019**	**0.945 ±0.014**	**0.910 ±0.015**	**0.935 ±0.021**	**0.905 ±0.028**	**0.934 ±0.019**

Table 8
Comparison of TSFM-Net and the ablation networks on CT images (Dice±Standard Deviation).

Method	LV	RV	LA	RA	LV_Myo	AA	PA	Mean
Backbone	0.908 ±0.051	0.911 ±0.032	0.919 ±0.031	0.886 ±0.039	0.874 ±0.029	0.933 ±0.024	0.854 ±0.056	0.898 ±0.037
Backbone+A	0.939 ±0.036	0.907 ±0.047	0.930 ±0.056	0.888 ±0.079	0.906 ±0.048	0.942 ±0.082	0.836 ±0.188	0.906 ±0.077
Backbone+A+B (TSFM-Net)	**0.950 ±0.025**	**0.936 ±0.026**	**0.951 ±0.022**	**0.922 ±0.034**	**0.926 ±0.014**	**0.973 ±0.012**	**0.922 ±0.034**	**0.940 ±0.019**

Table 9
Comparison of TSFM-Net and the ablation networks on MRI images (Dice±Standard Deviation).

Method	LV	RV	LA	RA	LV_Myo	AA	PA	Mean
Backbone	0.928 ±0.024	0.899 ±0.039	0.891 ±0.035	0.904 ±0.029	0.851 ±0.025	0.875 ±0.030	0.826 ±0.046	0.882 ±0.033
Backbone+A	0.956 ±0.014	0.938 ±0.024	0.935 ±0.022	0.940 ±0.016	0.904 ±0.019	0.931 ±0.016	0.897 ±0.030	0.928 ±0.020
Backbone+A+B (TSFM-Net)	**0.959 ±0.012**	**0.943 ±0.023**	**0.940 ±0.019**	**0.945 ±0.014**	**0.910 ±0.015**	**0.935 ±0.021**	**0.905 ±0.028**	**0.934 ±0.019**

Table 10
Balance evaluation of segmentation performance (Standard Deviation).

Method	CT	MRI
Backbone	0.028	0.035
Backbone+A	0.037	0.021
Backbone+A+B	**0.019**	**0.019**

ablation studies. We design a feature aggregation module (FAM) in the encoder to extract multi-scale features. Based on the diversity of features contained in the intermediate feature maps, we propose a multi-level attention mechanism (MLAM) for bilateral feature maps to pay more attention to targets, so as to reduce the restriction of single scale feature maps in segmentation guidance. As for these structural improvements, we demonstrate their effectiveness through the following ablation experiments.

Table 8 and 9 show the segmentation results of the ablation networks. The proposed segmentation network is based on the backbone segmentation network (Backbone), with
adding two modules of FAM (A) and MLAM (B), called Backbone+A+B (TSFM-Net). Therefore, we compare TSFM-Net with Backbone and Backbone+A.

As can be seen from the quantitative segmentation results in Table 8 and 9, our method achieves the best segmentation results on the 7 targets (including the mean value) for CT and MRI images. Figure 9 randomly visualizes the segmentation results of our method and the two ablation methods. All of these methods could achieve good segmentation performance for large-size targets. However, when the segmentation targets are scattered, TSFM-Net is able to better capture the boundary regions, while the ablation methods cannot. This advantage can be well reflected in the images of the second row. Thus, TSFM-Net could improve the balance of multiple targets to some extent by using multi-level attention mechanism, and pay more attention to small targets and scattered targets. The balance evaluation of segmentation performance is shown in Table 10.

In order to prove that the proposed method has better segmentation ability compared with the ablation methods, ROC curve and AUC value are adopted for further analysis. As shown in Figure 10, we randomly select the predicted probability values of several test images and their real category values to calculate the AUC values and draw the ROC curve. As can be seen from the figure, the proposed model has higher AUC values than the other two ablation models, especially in the segmentation of CT images.

In the whole segmentation framework, we make a detailed design, which is attention fusion module (AFM). To explore the impact of this detailed design on overall segmentation performance, we adopt the method of removing this module to verify its effectiveness. As shown in Table 11, "w/o" represents the removal of this module and "w/" represents the retention of this module. As can be seen from the quantization segmentation results, the role of AFM is prominent for CT images.

4.4.5. Visualization of attention maps

When facing with multi-target segmentation task, the homogeneity among multiple targets seriously affect the judgment of target category. In particular, the slices as network input lead to a serious imbalance in the number of samples, which directly affects the weight bias of the network in the training process. Thus, we propose a multi-level attention mechanism (MLAM), in attempt to improve this problem and increase the diversity of attention maps on scale. Through the action of MLAM, the network pays more attention to the target regions of each channel and treats each target channel equally. As shown in Figure 11, the input and output of MLAM are visualized by feature maps. For the three resolution feature maps, the attention mechanism strengthens the target regions or weakens the non-target regions to some extent. By doing so, the multi-level attention maps we obtained highlights the target areas on the whole.

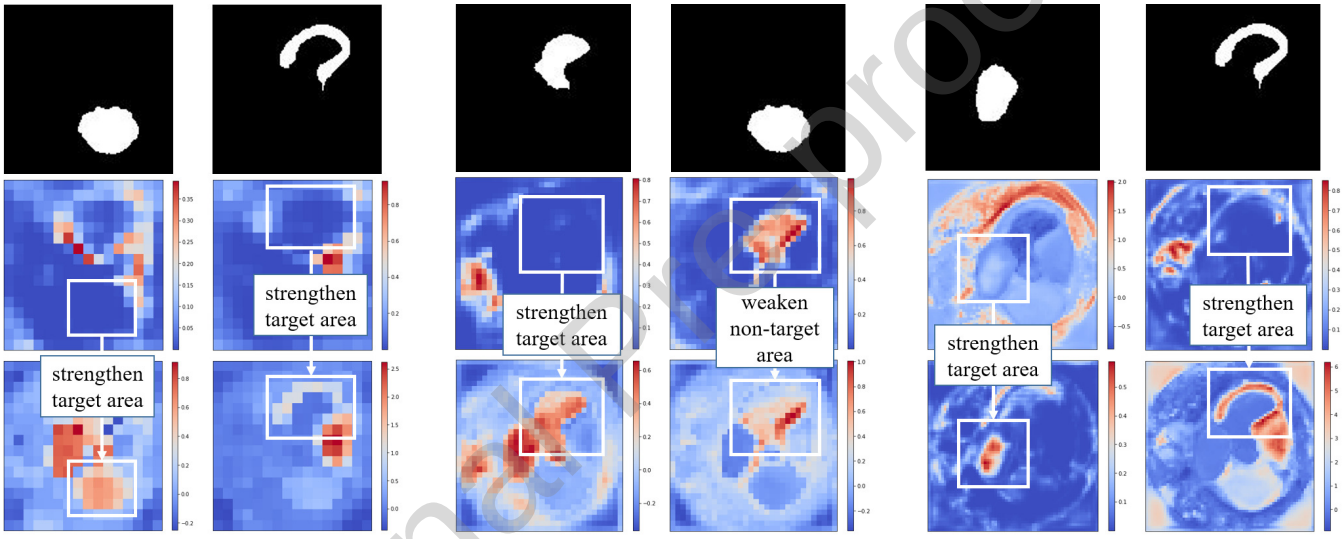
5. Conclusion

In this paper, a two-stage segmentation network with feature aggregation and multi-level attention mechanism (TSFM-Net) is proposed for multi-modality whole heart image segmentation. For $Stage1$ segmentation network, we propose a feature aggregation module (FAM) to supplement the traditional encoder. When extracting multi-level feature information, the feature validity can be improved by fusing the advanced feature maps of different levels. Through the training of $Stage1$ segmentation network, we have obtained the bilateral feature maps and sent them into a multi-level attention mechanism (MLAM), which contains three bilateral attention modules (BAM). In BAM, the globalmaxpooling layer and the globalaveragepooling layer are used to compress the multi-level feature maps into feature vectors, and make the significant parts between them fully interactive. The output of $Stage1$ is fused with the multi-level attention maps to form the segmentation masks of $Stage2$ segmentation network.

To verify the effectiveness of the proposed TSFM-Net for multi-target image segmentation task, TSFM-Net is applied to whole heart CT and MRI images to segment 7 cardiac substructures. A large number of comparative experiments and ablation studies have proved that TSFM-Net is more suitable for multi-target image segmentation than the existing methods. Based on the effective extraction of multi-level features by FAM, the more attention to multiple targets by MLAM and the two-stage segmentation strategy, our method not only achieve extremely competitive segmentation results, but also ensure that the segmentation performance of multiple targets can be maintained at the same level.

Table 11
Evaluation of the effect of AFM.

AFM	LV	RV	LA	RA	LV_Myo	AA	PA	Mean
				CT				
w/o	0.945 ±0.027	0.918 ±0.054	0.945 ±0.025	0.910 ±0.046	0.919 ±0.029	0.968 ±0.013	0.910 ±0.046	0.931 ±0.034
w/	**0.950 ±0.025**	**0.936 ±0.026**	**0.951 ±0.022**	**0.922 ±0.034**	**0.926 ±0.014**	**0.973 ±0.012**	**0.922 ±0.034**	**0.940 ±0.019**
				MRI				
w/o	0.958 ±0.014	0.942 ±0.023	0.939 ±0.022	0.944 ±0.014	**0.913 ±0.017**	**0.939 ±0.014**	0.902 ±0.028	**0.934 ±0.019**
w/	**0.959 ±0.012**	**0.943 ±0.023**	**0.940 ±0.019**	**0.945 ±0.014**	0.910 ±0.015	0.935 ±0.021	**0.905 ±0.028**	**0.934 ±0.019**



(a) Feature map resolution: 16×16 (b) Feature map resolution: 32×32 (c) Feature map resolution: 64×64

Figure 11: Visualization of multi-level attention maps.

6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

7. Acknowledgments

This work was supported in part by the Provincial Natural Science Research Program of Higher Education Institutions of Anhui province under grant KJ2020A0035. The authors thank all the anonymous reviewers for their valuable comments and suggestions, which were helpful for improving the quality of the paper. And then, the authors also acknowledge the High-performance Computing Platform of Anhui University for providing computing resources.

References

[1] Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence .

[2] Boot, T., Irshad, H., 2020. Diagnostic assessment of deep learning algorithms for detection and segmentation of lesion in mammographic images, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 56–65.

[3] Chen, C., Dou, Q., Chen, H., Qin, J., Heng, P., 2020. Unsupervised bidirectional crossmodality adaptation via deeply synergistic image and feature alignment for medical image segmentation 39(7), 2494–2505.

[4] Chen, S., Bortsova, G., Juarez, G.U., Tulder, G.V., Bruijne, M., 2019. Multi-task attention-based semi-supervised learning for medical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI).

[5] Du, X., Song, Y., Liu, Y., Zhang, Y., Li, S., 2020. An integrated deep learning framework for joint segmentation of blood pool and myocardium. Medical Image Analysis 62(101685).

[6] Fang, C., Li, G., Pan, C., Li, Y., Yu, Y., 2019. Globally guided progressive fusion network for 3d pancreas segmentation, in: Medical

Image Computing and Computer-Assisted Intervention (MICCAI), pp. 210–218.

[7] Fei, L., Kwab, C., Dan, L.D., Xin, Y., Jie, T., 2020. Deep pyramid local attention neural network for cardiac structure segmentation in two-dimensional echocardiography. Medical Image Analysis 67(101873).

[8] Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H., 2018. Dual attention network for scene segmentation .

[9] Galisot, G., Brouard, T., Ramel, J.Y., 2018. Local probabilistic atlases and a posteriori correction for the segmentation of heart images, Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges. pp. 207–214.

[10] Gao, Y., Liu, C., Zhao, L., 2019. Multi-resolution path cnn with deep supervision for intervertebral disc localization and segmentation, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 309–317.

[11] Gu, P., Zheng, H., Zhang, Y., Wang, C., Chen, D.Z., 2021. kcbac-net: Deeply supervised complete bipartite networks with asymmetric convolutions for medical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI).

[12] He, K., Zhang, X., Ren, S., Sun, J., 2015. Resnet50: Deep residual learning for image recognition .

[13] Heinrich, M.P., Oster, J., 2018. Mri whole heart segmentation using discrete nonlinear registration and fast non-local fusion, in: Statistical Atlases and Computational Models of the Heart ACDC and MMWHS Challenges, pp. 233–241.

[14] Kang, D., Woo, J., Kuo, C., Slomka, P., Dey, D., Germano, G., 2012. Heart chambers and whole heart segmentation techniques: a review. J ELECTRON IMAGING 21(010901).

[15] Li, H., Xiong, P., An, J., Wang, L., 2018. Pyramid attention network for semantic segmentation .

[16] Long, J., Shelhamer, E., Darrell, T., 2014. Fully convolutional networks for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 39(4), 640–651.

[17] Luo, G., Ran, A., Wang, K., Dong, S., Zhang, H., 2016. A deep learning network for right ventricle segmentation in short-axis mri, in: Computing in Cardiology Conference, pp. 485–488.

[18] Ma, C., Ji, Z., Gao, M., 2019. Neural style transfer improves 3d cardiovascular mr image segmentation on inconsistent data. Medical Image Computing and Computer-Assisted Intervention (MICCAI) 11765, 128–136.

[19] Mahendra, K., Varghese, A., Ganapathy, K., 2019. Fully convolutional multi-scale residual densenets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers. Medical Image Analysis 51, 21–45.

[20] Mathai, T.S., Gorantla, V., Galeotti, J., 2019. Segmentation of vessels in ultra high frequency ultrasound sequences using contextual memory, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 173–181.

[21] Mendis, S., Puska, P., Norrving, B., 2011. Global atlas on cardiovascular disease prevention and control: World health organization .

[22] Mortazi, A., Burt, J., Bagci, U., 2017. Multi-planar deep segmentation networks for cardiac substructures from mri and ct, in: Statistical Atlases and Computational Models of the Heart ACDC and MMWHS Challenges, pp. 199–206.

[23] Pace, D., Dalca, A., Geva, T., Powell, A., Hedjazi, M.M., Golland, P., 2015. Interactive whole-heart segmentation in congenital heart disease, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 80–88.

[24] Payer, C., SternHorst, a., Urschler, B., 2018. Multi-label whole heart segmentation using cnns and anatomical label configurations, in: Statistical Atlases and Computational Models of the Heart ACDC and MMWHS Challenges, pp. 190–198.

[25] Poudel, R.P.K., Lamata, P., Montana, G., 2016. Recurrent fully convolutional neural networks for multi-slice mri cardiac segmentation, in: Reconstruction, Segmentation, and Analysis of Medical Images, pp. 83–94.

[26] Qin, C., Bai, W., Schlemper, J., Petersen, S.E., Piechnik, S.K., Neubauer, S., Rueckert, D., 2018. Joint learning of motion estimation and segmentation for cardiac mr image sequences, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 472–480.

[27] Rad, R.M., Saeedi, P., Au, J., Havelock, J., 2020. Trophectoderm segmentation in human embryo images via inceptioned u-net. Medical Image Analysis 62(101612).

[28] Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation .

[29] Savioli, N., Vieira, M.S., Lamata, P., Montana, G., 2018. A generative adversarial model for right ventricle segmentation .

[30] Shi, Z., Zeng, G., Le, Z., Zhuang, X., Li, L., Yang, G., Zheng, G., 2018. Bayesian voxdrn: A probabilistic deep voxelwise dilated residual network for whole heart segmentation from 3d mr images, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 569–577.

[31] Wang, C., Macgillivray, T., Macnaught, G., Yang, G., D., N., 2018. A two-stage 3d unet framework for multi-class segmentation on full resolution image, in: Statistical Atlases and Computational Models of the Heart ACDC and MMWHS Challenges.

[32] Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., Tang, X., 2017. Residual attention network for image classification, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[33] Wang, G., Shapey, J., Li, W., Dorent, R., Demitriadis, A., Bisdas, S., Paddick, I., Bradford, R., Zhang, S., Ourselin, S., Vercauteren, T., 2019. Automatic segmentation of vestibular schwannoma from t2-weighted mri by deep spatial attention with hardness-weighted loss, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 264–272.

[34] Wang, H., Wang, S., Qin, Z., Zhang, Y., Xia, Y., 2021. Triple attention learning for classification of 14 thoracic diseases using chest radiography. Medical Image Analysis 67(3).

[35] Xia, W., Fortin, M., Ahn, J., Rivaz, H., Battié, M., Peters, T.M., Xiao, Y., 2019. Automatic paraspinal muscle segmentation in patients with lumbar pathology using deep convolutional neural network, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 318–325.

[36] Xing, X., Yuan, Y., Meng, Q.H., 2020. Zoom in lesions for better diagnosis: Attention guided deformation network for wce image classification. IEEE Transactions on Medical Imaging , 99.

[37] Xu, C., Howey, J., Ohorodnyk, P., Roth, M., Li, S., 2019. Segmentation and quantification of infarction without contrast agents via spatiotemporal generative adversarial learning. Medical Image Analysis 59(101568).

[38] Xu, C., Xu, L., Gao, Z., Shen, Z., Zhang, H., Zhang, Y., Du, X., Shu, Z., Dhanjoo, G., Liu, H., 2018. Direct delineation of myocardial infarction without contrast agents using a joint motion feature learning architecture. Medical Image Analysis 50, 82–94.

[39] Xu, C., Xu, L., Ohorodnyk, P., Roth, M., Li, S., 2020. Contrast agent-free synthesis and segmentation of ischemic heart disease images using progressive sequential causal gans. Medical Image Analysis 62(101668).

[40] Xue, W., Nachum, I.B., Pandey, S., Warrington, J., Leung, S., Li, S., 2017. Direct estimation of regional wall thicknesses via residual recurrent neural network 10265, 505–516.

[41] Yang, X., Bian, C., Yu, L., Dong, N., Heng, P.A., 2017. Hybrid loss guided convolutional networks for whole heart parsing, in: Statistical Atlases and Computational Models of the Heart ACDC and MMWHS Challenges, pp. 215–223.

[42] Yi, J., Jiang, M., Wu, P., Huang, Q., Metaxas, D.N., 2019. Attentive neural cell instance segmentation. Medical Image Analysis 55, 228–240.

[43] Zhang, J., Liu, M., Wang, L., Chen, S., Shen, D., 2019. Context-guided fully convolutional networks for joint craniomaxillofacial bone segmentation and landmark digitization. Medical Image Analysis 60(101621).

[44] Zhao, N., Tong, N., Dan, R., Sheng, K., 2019. Fully automated pancreas segmentation with two-stage 3d convolutional neural networks,

in: Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 201–209.

[45] Zheng, H., Yang, L., Han, J., Zhang, Y., Chen, D.Z., 2019. Hfa-net: 3d cardiovascular image segmentation with asymmetrical pooling and content-aware fusion, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 759–767.

[46] Zhuang, X., Shen, J., 2016. Multi-scale patch and multi-modality atlases for whole heart segmentation of mri. Medical Image Analysis 31, 77–87.