Computers, Materials & Continua CMC, vol.63, no.2, pp.995-1012, 2020

Multi-Task Learning Using Attention-Based Convolutional Encoder-Decoder for Dilated Cardiomyopathy CMR Segmentation and Classification

Chao Luo[1], Canghong Shi[1], Xiaojie Li[1,*], Xin Wang[4], Yucheng Chen[3], Dongrui Gao[1], Youbing Yin[4], Qi Song[4], Xi Wu[1] and Jiliu Zhou[1]

Abstract: Myocardial segmentation and classification play a major role in the diagnosis of cardiovascular disease. Dilated Cardiomyopathy (DCM) is a kind of common chronic and life-threatening cardiopathy. Early diagnostics significantly increases the chances of correct treatment and survival. However, accurate and rapid diagnosis of DCM is still challenge due to high variability of cardiac structure, low contrast cardiac magnetic resonance (CMR) images, and intrinsic noise in synthetic CMR images caused by motion artifact and cardiac dynamics. Moreover, visual assessment and empirical evaluation are widely used in routine clinical diagnosis, but they are subject to high inter-observer variability and are both subjective and non-reproducible. To solve this problem, we proposed an effective unified multi-task framework for dilated cardiomyopathy CMR segmentation and classification simultaneously, and we firstly update one independent encoder from both recovery decoder and parallel attention path sharing some partial weights. This can encode both task choices into good embedding, but each one can achieve significant improvements respectively from the given embedding. It consists of three branches: extraction path, attention path, and recovery path, which allows the model to learn more higher-level intermediate representations and makes a more accurate prediction. We validated our approach on a DCM dataset, which contains 1155 CMR LGE images. Experimental results show that our multi-task network has achieved accuracy of 97.63%, AUC of 98.32%, demonstrating effectively segmenting the myocardium, quickly and accurately diagnosing the presence or absence of dilation.

Keywords: Dilated cardiomyopathy, multitasking network, attention, classification, segmentation.

[1] College of Computer Science, Chengdu University of Information Technology, Chengdu, 610000, China.
[2] College of Information Science and Technology, Southwest Jiaotong University, Chengdu, 610000, China.
[3] West China Hospital, Sichuan University, Chengdu, 610000, China.
[4] AI Institute, CuraCloud Corporation, Seattle, 98101, USA.
[*] Corresponding Author: Xiaojie Li. Email: lixiaojie000000@163.com.
Received: 16 July 2019; Accepted: 29 August 2019.

1 Introduction

Dilated cardiomyopathy (DCM), characterized by dilatation and impaired contraction of the left or both ventricles, is a primary cardiomyopathy of unknown cause. It has become a leading cause of heart failure and heart transplantation in younger adults, or sudden cardiac death (SCD) [Caforio, Bottaro and Iliceto (2012); Cazeau, Ritter, Bakdach et al. (2010); Jefferies and Towbin (2010)]. Thus, immediate diagnosis and accurate assessment of dilated cardiomyopathy are critical for treatment and later recovery [Weekes, Wheeler, Yan et al. (2010); Hershberger, Morales and Siegfried (2010); Westenberg, Rj, Lamb et al. (2005)]. In routine clinical diagnosis, cardiac magnetic resonance (CMR), as a set of non-invasive magnetic resonance imaging (MRI) techniques, is widely used which would produce detailed pictures of the structures within and around the heart. It can not only display the anatomical changes of each chamber, large blood vessels and valves, but can also perform quantitative atrial and ventricular analysis for qualitative and semi-quantitative diagnosis. Moreover, CMR with late gadolinium enhancement (LGE) has emerged as the gold standard for infarct sizing and assessment of viability after myocardial infarction [Schelbert, Hsu and Anderson (2010)]. LGE has the advantages of multiple charts, and high spatial resolution allowing visualization of the whole heart and lesions, and its relationship with surrounding structures [Bauer, Wiest, Nolte et al. (2013); Haribabu, Bindu and Prasad (2012); Wang, Li and Liu (2013); Yin, Duan, Chu et al. (2016)]. This can help doctors study the structure and function of heart muscle, find the cause of a patient's heart failure or identify tissue damage due to a heart attack. Visual assessment and empirical evaluation are widely used in routine clinical diagnosis. However, immediate diagnosis and accurate assessment of patients with DCM remains difficult even for experienced cardiologists. Moreover, they are subject to high inter-observer variability, and the results are subjective and non-reproducible.

To solve the above problems, automatic computer-aided technologies are highly desirable. Many medical image segmentation and classification techniques have been developed to assist cardiologists to diagnose [Bar, Diamant, Wolf et al. (2015); Greenspan, Ginneken and Summers (2016); Moeskops, Wolterink, Velden et al. (2016)]. As the one of the most successful machine learning techniques today, convolutional neural network (CNN), has been been successfully applied to different image segmentation and classification tasks. For example, traditional CNN-based methods generally use image block around a given pixel as input for training and prediction of each pixel for segmentation. However, it has several disadvantages: high storage overhead, poor computational efficiency, and the size of the perceived area is limited by the size of the pixel block. Specially, the size of a pixel block is much smaller than the size of the entire image, and only some local features can be extracted, resulting in limited performance of the classification [Tao and Gao (2019)]. Fortunately, Jonathan Long et al. [Long, Shelhamer and Darrell (2014)] proposed the Fully Convolutional Networks (FCN) for image segmentation. It attempts to recover the category to which each pixel belongs from the abstract features. That is, the classification from the image level is further extended to the classification at the pixel level. Compared with traditional CNN-based image segmentation methods, FCN is more efficient by avoiding high storage overhead and can accept input images of any size [Wu, Zhang, Zhang et al. (2017)]. However, its segmentation results are not precise enough due to up-sampling and ignores the spatial regularization and lacking spatial consistency. To

overcome the shortcomings of FCN, Ronneberger and his team proposed U-Net for Biomedical Image Segmentation [Ronneberger, Fischer and Brox (2015)]. However, these methods can not directly apply on DCM segmentation.

Furthermore, deep learning techniques are also widely used in image classification field. A large number of networks has been proposed. The famous deep convolutional network VGG (e.g., VGG-16 and VGG-19) was proposed by Researchers at the Computer Vision Group at Oxford University and Google DeepMind. It has a significantly lower error rate than the previous state-of-the-art network architecture. However, VGG has a very significant drawback. When increase the depth of the network, lead to features loss and gradient dispersion. Therefore, to make up for the VGG shortcomings, He et al. [He, Zhang, Ren et al. (2016)] proposed a ResNet network which can effectively solve the problem of feature loss and gradient dispersion by constructing a structure of jump connections. Despite this, due to the high noise and low resolution in DCM medical images, ResNet cannot be directly applied to medical image classification tasks.

However, immediate diagnosis and accurate assessment of dilated cardiomyopathy remains a challenging problem. Moreover, single classification or segmentation task learning ignores the relationship between tasks and does not effectively use the inextricable links between classification or segmentation task. In this work, we aim to propose a unified multi-task framework for dilated cardiomyopathy CMR segmentation and classification simultaneously. From a pixel perspective, the segmentation task can be considered as a pixel-level classification task. In deep neural networks, parameters sharing can be done for multi-task learning. By sharing representations between related tasks, we can improve our network generalization on our original task. Therefore, to compensate for the shortcomings of classification and segmentation tasks, we design an attention-based convolutional encoder-decoder network for DCM segmentation and classification simultaneously. In the proposed method, we firstly update one independent encoder from both recovery decoder and parallel attention path sharing some partial weights. This can encode both task choices into good embedding, but each one can achieve significant improvements respectively from the given embedding. It consists of three branches: extraction path, attention path, and recovery path, which allows the model to learn more higher-level intermediate representations and makes a more accurate prediction. We have fully extracted the semantic features of the image abstraction and fully fuses the different scale features of the image. Specially, inside each Attention Module, a bottom-up top-down structure is used to unfold the feed-forward and feed-back attention process into a single feed-forward process, suppressing other irrelevant features. We validated our approach on a DCM dataset, which contains 1155 CMR LGE images. Experimental results have demonstrated the effectiveness of the proposed method.

2 Related work

2.1 U-Net

U-Nets yielded better image segmentation in medical imaging due to several main advantages. First, U-Net is full-convolution network which replace fully connected layer with convolutional layer. Since the fully connected layer must have a fixed image size that is not required by convolution, U-Net can input images of any size. Therefore, U-Net

is an end-to-end network. Second, U-Net fuse the features of each upsampling layer with the features of the corresponding downsampling layer. Upsampling can complement the features of the image, but it can cause feature loss problems, so it needs to be integrated with the features of the downsampling layer. However, it is not possible to use U-Net directly for myocardial segmentation tasks. It has been clearly identified: 1) The accuracy of the receptive field and positioning cannot be enhanced simultaneously. When the receptive field selection is comparatively larger, the dimension of the corresponding pool layer is expected to increase, it is expected to lead to lower positioning accuracy. Likewise, in the case of a small receptive field, the positioning accuracy is also expected to decrease. 2) A lot of noise exists in the myocardial image. The contrast between the myocardium and the surrounding tissue is quite poor, substantially lowering the accuracy of the segmentation. 3) The redundancy of using U-net for image segmentation tasks is excessively large. Owing to the fact that each pixel takes a patch, the similarity of patches of two adjacent pixels is quite high, resulting into an extensive amount of redundancy, making the network training very slow.

2.2 Multi-task learning

Compared with single-task learning, multi-task learning is a subfield of machine learning in which multiple related tasks are learned simultaneously by transferring knowledge between tasks to improve the learning process. As shown in Fig. 1, single-task learning ignores the relationship between tasks, but in actual learning tasks, there are always inextricable relationship between different tasks. The advantage of multitasking is that it can take advantage of the internal relationships between different tasks, it mainly includes 2 points: 1) Weight sharing of neurons of different tasks. In general, some seemingly different tasks actually have intrinsic links. For example, the semantic segmentation of an image is actually a pixel-level classification. 2) Mutual promotion between different tasks. For deep learning, massive amounts of data are critical, however, getting massive amounts of data is very difficult. From the perspective of data enhancement, multi-task learning is equivalent to adding random noise to another task, solving the problem of a small number of data sets to a certain extent, and improving the generalization performance and robustness of the network.

Due to the above advantages of multitasking, it has been widely used in tasks such as image processing and natural language processing. Microsoft's Liu et al. proposed a new multi-tasking deep neural network model MT-DNN. MT-DNN combines the advantages of BERT, effectively utilizing supervised data from many related tasks, and benefiting from regularization effects by mitigating over-fitting of specific tasks, thereby enabling learning weight to be embedded between tasks. The network has created new and most advanced results in multiple mainstream benchmarks. Xue et al. use multitasking learning methods to accomplish different classification tasks. They greatly improve the performance of classification by sharing the weights of neurons of different classification tasks and learning the deep intrinsic association features between different tasks. The experimental results show that the classification accuracy of the multi-task method is much higher than the accuracy of the single classification task. Collobert and his team proposed a natural language processing model based on multitasking learning. The model

shares the weights and parameters of the supervised learning network and the semi-supervised learning network. Compared with the traditional natural language processing network, the multi-tasking network can accurately perform tasks such as language role judgment, semantic similar sentence discrimination, and voice tagging.

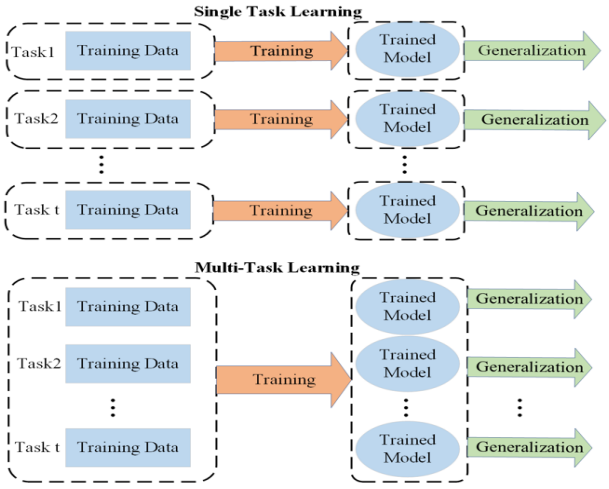

Figure 1: Model comparison diagrams of multi-task method and single-task method

2.3 Attention mechanism

Attention in deep learning originates from the attention mechanism of the human brain. When the brain receives external information, such as visual information and auditory information, it will not process and understand all information, but will focus on some significant or interesting portions. This will help to filter out the unimportant information and improve the efficiency of information processing. Fei Wang et al. proposed the residual attention network, which can greatly improve the accuracy of target classification by drawing on the attention mechanism.

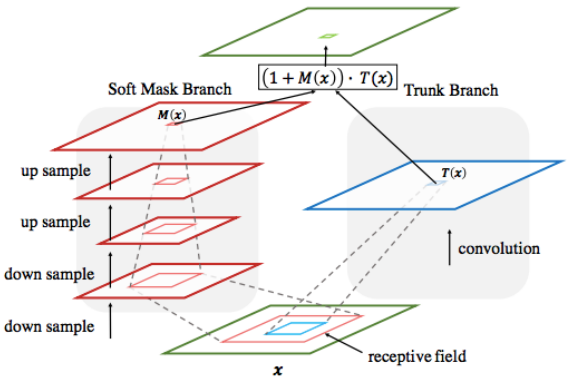

Figure 2: Attention model

There are currently many attention-based methods. The main contribution of our paper is

to propose an Attention Module, which uses the forward attention mechanism of Bottom-up Top-down. As shown in Fig. 2, the module consists of trunk branch and soft mask branch. The former is composed of multi-layer convolution, and first extracts high-level features and enlarges the receptive field of the model. Simultaneously, the latter enlarges the size of the feature map to the same size as the original input through the same number of up samples, so that the area of attention corresponds to each pixel of the input. Finally, the Soft Mask Branch is integrated with the output of the Trunk Branch. Each pixel value in the attention map output of the Soft Mask Branch is equivalent to the weight of each pixel value on the original feature map, this enhances the meaningful features and suppresses no meaning information. Therefore, as shown in Eq. (1), the element-wise multiplication of the feature map output by Soft Mask Branch and Trunk Branch gives a weighted attention map. The weighted attention map performs an element-wise operation with the feature map of the original Trunk Branch.

$$H_{i,c}(x)=(1+M_{i,c}(x))\times T_{i,c}(x) \tag{1}$$

where i ranges over all spatial positions, c is the index of the channel, $M(x)$ is the output of Soft Mask Branch and $T(x)$ is the output of the Trunk Branch.

2.4 Residual neural network

Residual neural network (ResNet) was proposed by He et al. [He, Zhang and Ren et al. (2016)]. ResNet with 152-layer neural network was successfully trained and won the championship in the ILSVRC2015 competition. The error rate on top 5 was 3.57%, and the parameter amount was lower than that of VGGNet. The main idea of ResNet is to add a direct connection channel to the network. It is very similar to that of Highway Network, which allows the original input information to be transmitted directly to the later layer.

We know that it is difficult to fit a potential identity mapping function $H(x)=x$ directly at the network level, which is the reason why deep networks are difficult to train. However, if the network is designed as $H(x)=F(x)+x$, it can be converted into learning a residual function $F(x)=H(x)-x$. As shown in Fig. 3, as long as $F(x)=0$, it constitutes an identical mapping $H(x)=x$. Moreover, fitting residual must be easier.

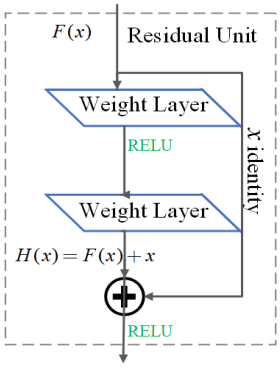

Figure 3: Residual unit

3 Method

In this paper, we propose a multitasking network structure based on the attention mechanism. The attention mechanism has the advantage of being able to highlight features that are advantageous for classification tasks and to suppress useless features that are not conducive to classification tasks. In addition, we construct an encoder-decoder structure for segmentation task and simultaneously learn the classification task and segmentation task. Multi-tasking can improve the generalization ability of the model and can also improve the network training efficiency to improve the performance of the network.

3.1 Network architecture

The residual module and the attention module are used as basic units in the construction of networks, our network architecture is shown in Fig. 4. It consists of an extraction path, an attention path, and a recovery path. The attention path is used for classification tasks and the recovery path is used for classification tasks. The extraction path follows the typical architecture of a convolutional network. It consists of seven 1×1 convolutions and three residual block, each convolution followed by a normalized layer and a rectified linear unit (RELU) and a 2×2 max pooling operation with stride 1 for downsampling. Every step in the recovery path consists of an upsampling of the feature map followed by a deconvolution that halves the number of feature channels, a summation with the correspondingly cropped feature map from the extraction path. The attention path consists with Soft Branch and Mask Branch, the module highlights features that are beneficial to the task and suppresses features that are not conducive to the task. The cropping is necessary because of the loss of border pixels in every convolution.

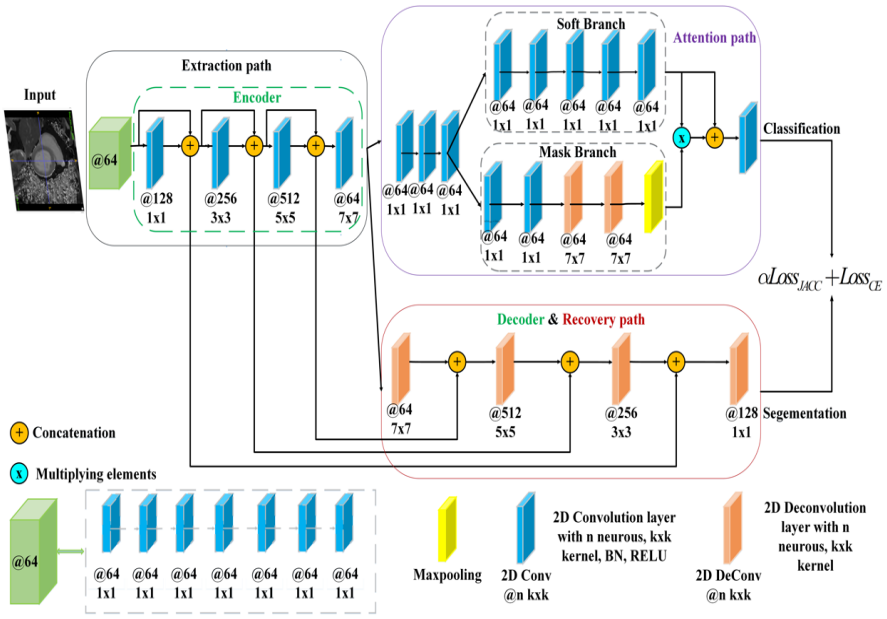

Figure 4: Network architecture

3.1.1 Extraction path

Extraction path consists of seven convolution layers with kernel size 1×1 and three residual units. It can extract high-level abstract features of input images and effectively avoids feature loss. Firstly, the image is input into seven convolution layers to get the high-level features of the image. Because the convolution layer will lose the features of the original data, the features obtained from the previous layer are input into three residual units to obtain the corresponding high-level features. Residual unit constructs a jump connection structure, which takes the input and output of the upper layer as the input of the next layer. It can effectively avoid feature loss, and also solve the problem of gradient disappearance caused by the increase of network layers. The calculation formula of the jump connection structure is as follows:

$$a^l = g(z^l + z^{l-1}) = g(w^l a^{l-1} + b^l + a^{l-1}) = g((w^l + 1)a^{l-1} + b^l) \tag{2}$$

l is the number of the current layer, a^l is the input of the current network layer, z^l is the output of the previous layer, a^{l-1} is the input of the previous layer, g is the activation function (the RELU we use in this paper), w^l is the weight of the current network layer neurons, b^l is the paranoia of the current network layer. The input a^l of the current network layer is equal to the sum of the output z^l of the upper layer network and the input z^{l-1} of the upper layer network, and then the activation function is connected.

3.1.2 Attention path

The attention path consists of two Attention Modules, constructed in reference to the residual attention network. The Attention Module consists of two branches, a soft branch and a mask branch. The soft branch consists of multiple convolution layers, and the mask branch consists of the encoder layer and the decoder layer. In the soft branch, we construct multiple convolutions to learn the deep features. The number of convolutional layers can be set according to the experimental results. In this paper, we set the number of convolutional layers to 4. In the mask branch, we construct the encoder-decoder structure. Two convolution kernels with a convolution kernel size of 1×1 are used for the encoder, and two deconvolution kernels with a convolution kernel size of 7×7 are used for the decoder. The characteristics of the classification can be learned through the encoder-decoder structure, and convolution kernels of different sizes can learn the characteristics of different fields of view. However, since the strategy of supplement 0 in deconvolution will lead to the increase of useless and unfavorable features and feature loss, we fuse the features of mask branch output with those of soft branch output, which can effectively compensate for the problem of feature noise and feature loss. For the process of attention path, first, the output of the soft branch is multiplied by the output of the mask branch, which filters out features that are not conducive to classification. Then, the output of the previous step is added to the output of the soft branch, avoiding the loss of features in the original data that are advantageous for classification. This allows the module to acquire features that are beneficial to classification and to suppress features that are not conducive to classification.

3.1.3 Recovery path

The recovery path consists of four different channel sizes of deconvolution layers. The strategy of supplementing 0 used by the deconvolution layer is therefore likely to cause feature loss and feature dispersion problems. In order to solve the above problems, we received the U-Net inspiration, we constructed a jump connection structure, that is, the output of each deconvolution layer is merged with the output of the convolution layer of the previous residual unit, and input to the next In a deconvolution layer. Each deconvolution layer adopts a jump connection structure, which effectively avoids the problem of feature dispersion and feature loss, and greatly improves the feature reuse rate. Recovery path can restore the size of the extracted deep features to the same size of the input image, that is, inputting an image can directly obtain a predicted segmentation image of the same size as the input image.

3.2 Loss metric

3.2.1 Loss function for segmentation

An inevitable difficulty in medical image classification and segmentation tasks is the class imbalance problem. The number of background pixels is much larger than the target pixel. Therefore, we need to choose a loss function that is immune to class imbalance problems. In this paper, we use the Jaccard (JACC) distance function as the loss function of the segmentation task, and the cross entropy (CE) function as the loss function of the classification task. The Jaccard index, also known as Intersection over Union (IoU), is a statistic used for assessing segmentation performance when ground-truth is available, as in Eq. (3):

$$Loss_{JACC} = 1 - \frac{|G \cap Y|}{|G \cup Y|} = 1 - \frac{|G \cap Y|}{|G| + |Y| - |G \cap Y|} = 1 - \frac{\sum_i^N G_i Y_i}{\sum_i^N G_i + \sum_i^N Y_i - \sum_i^N G_i Y_i} \tag{3}$$

where G is the ground-truth label map, $G_i \in (0,1)$. Y is the predicted probabilistic response map by model, $Y_i \in (0,1)$. N is the pixel number.

3.2.2 Loss function for classification

Cross entropy is often used for the classification loss function of deep neural networks. Here we use cross entropy as the loss function of the myocardial classification. Formally, it is defined as:

$$Loss_{CE} = -\frac{1}{m}\sum_i^m [y^i \log p^i + (1 - p^i)\log(1 - y^i)] \tag{4}$$

where m is the number of samples, y^i is the label of the sample, and p^i is the predicted probability value, $p^i \in (0,1)$.

In the experiment, the total loss $Loss_{total}$ function is the sum of the two loss function, but

because the segmentation task is relatively complex, the network needs to pay more attention to the segmentation task. So, we set the weight of the classification loss to 1, while constantly adjusting the weight of the segmentation loss to achieve the best performance. The total loss is defined as:

$$Loss_{total} = \partial Loss_{JACC} + Loss_{CE} \tag{5}$$

where $Loss_{JACC}$ is the loss function of the segmentation task, $Loss_{CE}$ is the loss function of the classification task, and ∂ is the weight of $Loss_{JACC}$. By changing the value of ∂, the network achieves the best performance. In addition, we did many experiments with different weights of segmentation loss. As shown in Fig. 5, results show that when the weight of segmentation loss is 3, the classification and segmentation performance is the best.

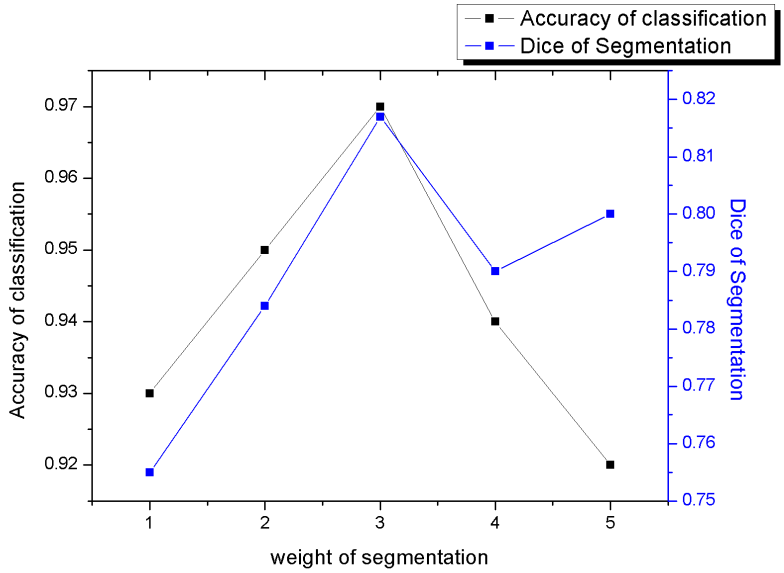

Figure 5: Segmentation and classification results corresponding to weights of different segmentation losses

4 Experiments and discussion

4.1 Dataset, pre-processing and evaluation metrics

4.1.1 Dataset

CMR LGE images of 1155 slices of myocardium from the same hospital are trained and tested in the experiments, which contain 123 dead images and 1032 alive images with heart disease. For the segmentation task, the original image was the cardiac MRI raw data, and the label represents the myocardium manually labeled by the doctor. For the classification task, we divide the LGE data into dead or alive.

4.1.2 Pre-processing

The ground truth of the training and the testing samples of images with the myocardium were manually annotated by an experienced cardiologist. The pixel size of each MRI image was 1.406×1.406 mm with a size ranging between 192×256 and 224×256. Considering the fact that the size of the myocardium is small and surrounded by a substantial amount of noise, we first resampled the resolution of the image to 1×1×1, and each image was cut to a fixed image size of 128×128. Finally, the intensity range of the LGE-mapping images was normalized to [0, 255].

4.1.3 Evaluation Metrics for Segmentation

We use of four indicators for the purpose of measure the performance of the network, which includes the dice similarity coefficient (DSC), area under the curve (AUC), Jaccard similarity coefficient (JSC), and F1-score for the assessment of the segmentation accuracy. The DSC was mostly employed for the calculation of the overlap metric between the results of segmentation and the ground truth. The DSC for bit vectors was defined as:

$$DSC = \frac{2 \| PG \|_2}{\| P \|_2 + \| G \|_2} \tag{6}$$

where PG is the element-wise product of the prediction P and the ground truth G, and $\| x \|_2$ is the L2-norm of x. The AUC is a probability value. The greater the AUC value, the better the performance. The AUC score was computed with a closed-form formula:

$$AUC = \frac{S_0 - n_0(n_0 + 1) \div 2}{n_0 n_1} \tag{7}$$

where n_0 is the number of pixel that belong to the ground truth, n_1 is the opposite and $S_0 = \sum_{i=1}^{n_0 r_i}$, where r_i is the rank given by the predict model of the ground truth to the i-th pixel in the CMR image. The F1 score is the harmonic average of precision and recall, wherein an F1 score reaches its best value at one (perfect precision and recall) and the worst at zero. The JSC is put to use for the improvement of similarities and differences between finite sample sets. The larger the JSC value, the higher the sample similarity.

4.1.4 Evaluation Metrics for classification

For the two-class task, we can define the discriminant evaluation based on the confusion matrix, as shown in Tab. 1. From the confusion matrix, TP and TN are defined as the number of positive and negative samples correctly classified. FP and FN are respectively defined as the number of positive and negative samples that are misclassified.

Table 1: Confusion matrix

	Actual Positive	Actual Negative
Predicted Positive	True Positive (TP)	False Negative (FN)
Predicted Negative	False Positive (FP)	True Negative (TN)

We used four indicators to assess the performance of the network, which includes accuracy, sensitivity, specificity and AUC. Accuracy is the ratio of the correct number to the total quantity, it is the most commonly used indicator for measuring classification results. Accuracy is computed using a closed-form formula:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{8}$$

Sensitivity is the probability of correctly predicting positive samples, and specificity is the probability of correctly predicting negative samples. Sensitivity is computed using a closed-form formula:

$$Sensitivity = \frac{TP}{TP + FN} \tag{9}$$

where TP is the number of positive samples that are correctly predicted and FN is the number of negative samples with incorrect prediction results. Similarly, specificity is computed using a closed-form formula:

$$Specificity = \frac{TN}{TN + FP} \tag{10}$$

where TN is the number of negative samples that are correctly predicted and FP is the number of positive samples with incorrect prediction results. Generally, the higher the above four indicators, the better the performance of the experimental method.

4.2 Implementation details

In our experiment, 5-fold cross validation was used for performance evaluation and method comparison. We calculated the average score of each fold as the final score. The network is implemented by Keras and Tensorflow backend. Batch normalization and ReLU activation are attached to each convolution. We use the Adam optimizer with a learning rate of 0.0001. The cost functions of segmentation and classification are Jaccard distance and cross-entropy, and the weight of loss function of segmentation task and classification task are 3 and 1, respectively. Model training and testing run on Tesla M40 GPU with 40 epochs.

4.3 Result

To verify the performance of our network, for the segmentation task, we compared it with the classic segmentation networks U-Net, DeeplabV3 and SegNet. For the classification task, we compared it with the Residual attention networks, ResNet50 and ResNet101. Experiments on

LGE datasets showed state-of-the-art performance. We implemented all networks on the basis of tensorflow, and parameters were set to be the same as that of our network.

4.3.1 Segmentation result

Tab. 2 shows a comparison of the results of our method with the most advanced methods available. As can be seen from the table, our method is further improved and achieves around 3% absolute improvement over the best previous method DeeplabV3 on DSC. Our method also outperforms previous methods on AUC, F1-score and JSC with a large margin. Surprisingly, U-Net and SegNet, which are the best-performing method on medical image segmentation, did not achieve good results on our dataset.

Table 2: Experimental results of different networks for segmentation task

Model	DSC	AUC	F1-score	JSC
DeepLabV3	0.7806	0.8561	0.7729	0.6127
SegNet	0.7496	0.7751	0.7187	0.5155
U-Net	0.7613	0.8168	0.7403	0.5611
Ours	**0.8165**	**0.9230**	**0.8164**	**0.6966**

Fig. 6 shows a block diagram of the DSC values for four networks. As can be seen from the figure, the span of our method block diagram was small, and the DSC values of the five experiments were mainly concentrated around \$0.81\$. However, the spans of U-Net, DeeplabV3 and SegNet were significantly larger than ours, and the DSC values of our five experiments were larger than those of U-Net, DeepLabV3 and SegNet. It can be seen that our method not only had better performance than U-Net, DeeplabV3 and SegNet, but also had better robustness.

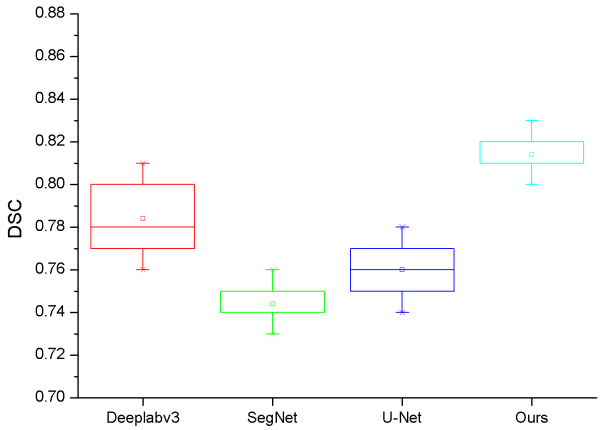

Figure 6: Box diagram of DSC for four different networks. Each network performs 5 experiments, and the box plot was drawn based on the DSC value of each experimental result

To quantificationally evaluate the difficulty of the myocardium segmentation task, we adopt the Peak Signal to Noise Ratio (PSNR) for image segmentation to describe. The PSRN is defined as the Eq. (11):

$$PSNR = 20\log_{10}(\frac{MAX_{ori}^2}{\sqrt{MSE}}) \tag{11}$$

where the MAX_{ori} is the max value of original image. And MSE is the Mean Square Error function, define as the Eq. (12):

$$MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[I_{ORI}(i,j) - I_{SEG}(i,j)]^2 \tag{12}$$

where I_{ORI} is the original image, T_{SEG} is the segmentation target, and the m, n are the image dimensions.

In Fig. 7, we calculated the peak signal-to-noise ratio (PSNR) values for 1155 LGE mapping images, which ranged from 11.60 to 31.15. Obviously, the PSNR values are mainly concentrated around 23.

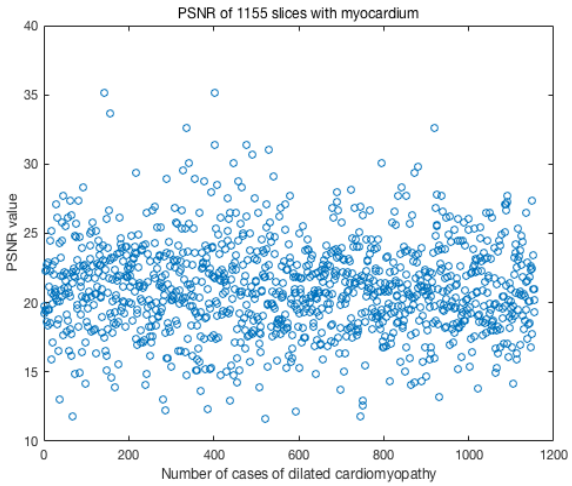

Figure 7: Scatter diagram PSNR of 1155 slices with myocardium

It may be observed from Fig. 8 that the proposed network is capable of segmenting the myocardial circle efficiently. Through the comparison of the segmentation results with U-Net, DeeplabV3 and SegNet, we found that the segmentation result of our network was not only less noisy but also closer to the ground truth. In particular, by comparing the segmentation results in Fig. 8, we can see that U-NET, Deeplabv3 and SegNet will generate a lot of noise to interfere with myocardial segmentation, and even cannot accurately and completely segment the myocardial region. However, our myocardial segmentation is fairly smooth and achieves good results in the case of strong noise.

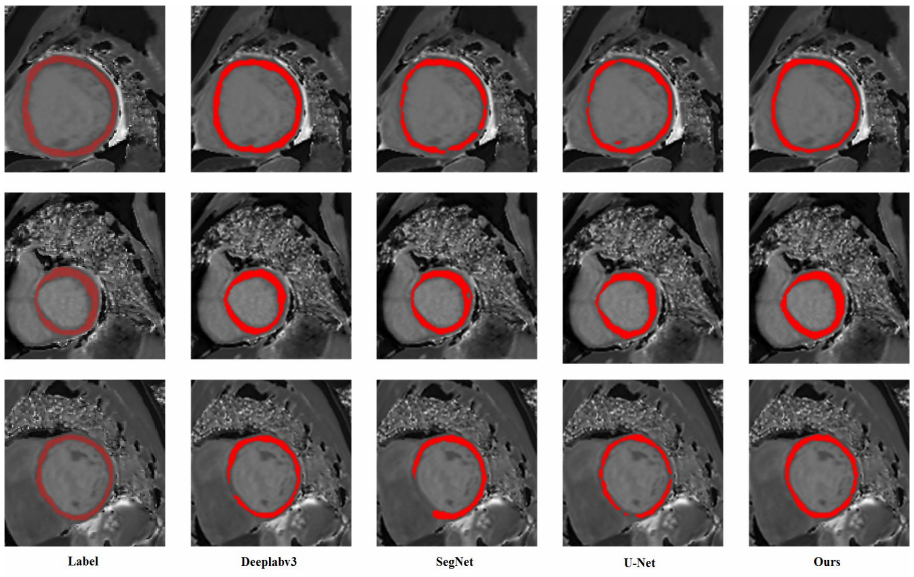

Figure 8: Comparison of segmentation results of different networks

4.3.2 Classification results

From Tab. 3 we can see that the classification performance of our method is much better than that of the other three networks on the LGE data set. Surprisingly, our method was 5% higher than the most advanced method ResNet101 in accuracy, and our method was far superior to ResNet101 in both sensitivity and specificity. In addition, the accuracy, specificity and sensitivity of our method are all the highest at the same time, which shows that our method has the best robustness.

Table 3: Experimental results of different networks for classification task

Model	Accuracy	Sensitivity	Specificity	AUC
Residual attention network	0.9258	0.1026	0.8375	0.9532
ResNet50	0.9208	0.2251	0.8261	0.9429
ResNet101	0.9431	0.2870	0.8729	0.9623
Ours	**0.9763**	**0.7592**	**0.9164**	**0.9832**

To further understand how the classifier predicts, we visualize the class-activation map (CAM) of the last convolutional layer. Negative and positive groups were shown in Fig. 9. There was a significant difference in the corresponding CAM of the negative (alive

control) and positive (death) samples. We can see that the heat map of the negative sample is more concentrated than the heat map of the positive sample.

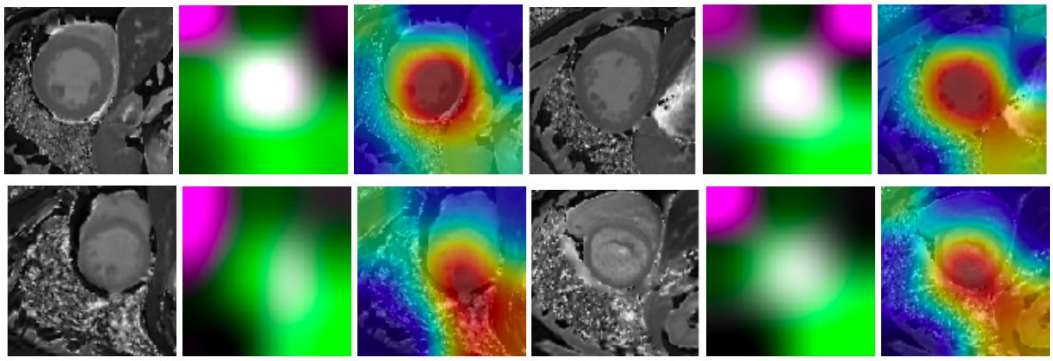

Figure 9: Class activation maps. The first row is examples of negative samples. The second row is examples of positive samples. (a) the original CMR images. (b) the activation maps. (c) The heatmap is overlaid on the original images

5 Conclusion

In this paper, we propose a muti-task approach for classification and segmentation. We demonstrate that our network trained using a multi-task method reaches the current state-of-the-art performance, making it much better than comparable methods. Experimental results show that the performance of multi-task segmentation and classification is better than that of single-task segmentation or single-task classification. In addition, we used the attention mechanism, which greatly improved the overall performance of the network. Multi-scale convolution has the ability to perceive multi-scale features of the myocardium as well as hierarchical semantics and contextual information. To prove the validity of our proposed network, we compared it with the classic segmentation network and classification network. Experiment reveals the fact that the proposed network can significantly improve segmentation and classification performance.

Acknowledgement: This work was supported by the National Natural Science Foundation of China (61602066), the Project of Sichuan Outstanding Young Scientific and Technological Talents (19JCQN0003), the major Project of Education Department in Sichuan (17ZA0063 and 2017JQ0030), and in part by the Natural Science Foundation for Young Scientists of CUIT (J201704) and the Sichuan Science and Technology Program (2019JDRC0077).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

Bar, Y.; Diamant, I.; Wolf, L.; Lieberman, S.; Konen, E. et al. (2015): Chest pathology detection using deep learning with non-medical training. *IEEE International Symposium on Biomedical Imaging*, pp. 294-297.

Bauer, S.; Wiest, R.; Nolte, L. P.; Reyes, M. (2013): A survey of MRI-based medical image analysis for brain tumor studies. *Physics in Medicine Biology*, vol. 58, no. 6, pp. R97.

Caforio, A. L.; Bottaro, S.; Iliceto, S. (2012): Dilated cardiomyopathy (DCM) and myocarditis: classification, clinical and autoimmune features. *Applied Cardiopulmonary Pathophysiology*, vol. 16, no. 1, pp. 82-95.

Cazeau, S.; Ritter, P.; Bakdach, S.; Lazarus, A.; Limousin, M. et al. (2010): Four chamber pacing in dilated cardiomyopathy. *Pacing Clin Electrophysiol*, vol. 17, no. 11, pp. 1974-1979.

Greenspan, H.; Ginneken, B. V.; Summers, R. M. (2016): Guest editorial deep learning in medical imaging: overview and future promise of an exciting new technique. *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1153-1159.

Hershberger, R. E.; Morales, A.; Siegfried, J. D. (2010): Clinical and genetic issues in dilated cardiomyopathy: a review for genetics professionals. *Genetics in Medicine*, vol. 12, no. 11, pp. 655-667.

Haribabu, M.; Bindu, C. H.; Prasad, K. S. (2012): Multimodal medical image fusion of MRI-pet using wavelet transform. *International Conference on Advances in Mobile Network*, pp. 127-130.

He, K.; Zhang, X.; Ren, S.; Sun, J. (2016): Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778.

Jefferies, J. L.; Towbin, J. A. (2010): Dilated cardiomyopathy. *Lancet*, vol. 375, no. 9716, pp. 752-762.

Long, J.; Shelhamer, E.; Darrell, T. (2014): Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis Machine Intelligence*, pp. 3431-3440.

Moeskops, P.; Wolterink, J. M.; Velden, B. H. M. V. D.; Gilhuijs, K. G. A.; Leiner, T. et al. (2016): Deep learning for multi-task medical image segmentation in multiple modalities. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 478-486.

Ronneberger, O.; Fischer, P.; Brox, T. (2015): U-Net: convolutional networks for biomedical image segmentation. *International Conference on Medical Image Computing and Computer-assisted Intervention*, pp. 234-241.

Schelbert, E. B.; Hsu, L. Y.; Anderson, S. A.; Mohanty, B. D.; Karim, S. M. et al. (2010): Late gadolinium-enhancement cardiac magnetic resonance identifies postinfarction myocardial fibrosis and the border zone at the near cellular level in *ex vivo* rat heart. *Circulation: Cardiovascular Imaging*, vol. 3, pp. 743-752.

Shen, T.; Nagai, Y.; Gao, Chan. (2019): Improve computer visualization of architecture based on the Bayesian network. *Computers, Materials & Continua*, vol. 58, no. 2, pp. 307-318.

Weekes, J.; Wheeler, C. H.; Yan, J. X.; Weil, J.; Eschenhagen, T. et al. (2010): Bovine dilated cardiomyopathy: proteomic analysis of an animal model of human dilated cardiomyopathy. *Electrophoresis*, vol. 20, no. 4, pp. 88-906.

Westenberg, J. J.; Rj, V. D. G.; Lamb, H. J.; Versteegh, M. I.; Braun, J. et al. (2005): MRI to evaluate left atrial and ventricular reverse remodeling after restrictive mitral annuloplasty in dilated cardiomyopathy. *Circulation*, vol. 112, no. 9, pp. 437-442.

Wang, X.; Li, W. L.; Liu, F. (2013): New algorithm of CT/MRI medical image fusion based on wavelet domain. *Journal of Jilin University*, vol. 43, no. S1, pp. 25-28.

Wu, Y.; Zhang, Y.; Zhang, C.; He, Z.; Zhang, Y. (2017): Semantic segmentation of mechanical parts based on fully convolutional network. *9th International Conference on Modelling, Identification and Control*, pp. 612-617.

Yin, M.; Duan, P.; Chu, B.; Liang, X.; Mathematics, S. O. (2016): CT and MRI medical image fusion based on shift-invariant shearlet transform and compressed sensing. *Opto-Electronic Engineering*, no. 8, pp. 8.