ACNET: Attention-based Convolution Network with Additional Discriminative Features for DCM Classification

Chao Luo
College of Computer Science
Chengdu University of
Information Technology
Chengdu, China
clchaoluo@163.com

Wang Xin
College of Computer Science
Chengdu University of
Information Technology
Chengdu, China
xinw@curacloudcorp.com

Xiaojie Li*
College of Computer Science
Chengdu University of
Information Technology
Chengdu, China
lixiaojie000000@163.com

Yucheng Chen*
West China Hospital
Sichuan University
Chengdu, China
chenyucheng2003@126.com

Jiliu Zhou
College of Computer Science
Chengdu University of
Information Technology
Chengdu, China
zhoujiliu@cuit.edu.cn

Kunlin Cao & Youbing Yin
CuraCloud Corporation
Seattle, USA
cao, yin@curacloudcorp.com

Qi Song
CuraCloud Corporation
Seattle, USA
song@curacloudcorp.com

Xi Wu
College of Computer Science
Chengdu University of
Information Technology
Chengdu, China
xi.wu@cuit.edu.cn

***Abstract*—For dilated cardiomyopathy (DCM) patients, immediate emergency diagnosis and treatment are critical for life saving and later recovery. T1 mapping is a non-invasive and effective diagnostic imaging approach to detect DCM. However, it is a demanding and time-consuming approach. In this paper, we propose an attention-based network structure, which can automatically identify DCM patients in a speedy manner to prioritize their treatment. In the proposed method, we adopt attention modules to generate attention-aware features. Inside each attention module, a bottom-up top-down feed-forward structure is used to unfold the feed-forward and feed-back attention processes into a single feed-forward process. It allows the network to focus more on determining useful information about the current output that is significant in the input data. Moreover, inspired by the residual network idea, we make full use of the characteristics of the original data. Combined residual block, we design down-residual modules for classification tasks. It consists of seven convolution layers and three layers of residual blocks. Our network achieves the most advanced recognition performance on cardiac datasets. We evaluated our approach on CMR(cardiac magnetic resonance) T1 mapping images with lower PSNR(peak signal to noise ratio), and the results demonstrate that our architecture outperforms previous approaches.**

***Index Terms*—medical image classification, attention, classification, myocardial**

I. Introduction

Image classification is one of the most important tasks in computer vision tasks [1], and it is also the basis of other high-level visual tasks such as object tracking and behavior analysis [2] [3]. Particularly in medical image tasks, medical image classification is the key issue to determining whether medical images can provide a reliable basis for clinical diagnosis and treatment [4]. The development of medical classification technology plays an extremely important role in biomedical image analysis. In recent years, classification technology has made significant progress because of the application of deep learning algorithms in medical image classification. This is more challenging than classification tasks on natural images. The highly cluttered background and the particularity of medical images pose major challenges to classification accuracy [5] [6].

Dilated cardiomyopathy (DCM) is a common myocardial disease [7] [8] [9]. The disease can lead to ventricular systolic dysfunction, congestive heart failure and arrhythmia. The disease is progressively aggravated, and death can occur at any stage of the disease. Therefore, the use of DCM data as a dataset is of great significance [10] [11] [12] [13].

At present, many methods use medical image classification, such as, the SVM-based(Support Vector Machine) classification method [14] [15]and texture-based method [16] [17]. However, on comparing various methods, the method based on an artificial neural network achieves better results. It is found that the artificial neural network has learning and recognition abilities similar to the human brain. Moreover, it can model human organs, and independently learn the determined

* These authors are co-corresponding authors. This work was supported by the National Natural Science Foundation of China (Grant No. 61602066) and the Scientific Research Foundation (KYTZ201608) of CUIT and the major Project of Education Department in Sichuan (17ZA0063 and 2017JQ0030), and partially supported by the Sichuan international science and technology cooperation and exchange research program (2016HH0018). DOI reference number: 10.18293/SEKE2019-155

diagnostic information and diseased tissue. Additionally, it can improve the reliability and effectiveness of diagnosing a patient's condition. Therefore, in this paper, we use an attention-based neural network method to effectively classify medical images.

Inspired by the residual attention network and recent advances in deep neural networks [18] [19], we propose a new network based on attention. It is composed of two attention modules that generate attention-aware features. Additionally, we design the down-residual module. It is capable of extracting the high-level features of the input data without causing a loss of the original features. As the number of layers increases, the attention-aware features from different modules are adaptively changed.

In addition to the additional discriminating features acquired by the attention module, our network has the following advantages: 1) Before the attention module, we design a down-residual module, which enables the extraction of high-level features of the input data without causing a loss of the original features. 2) Because of the scalability of the attention module, our network can be easily extended to hundreds of layers. Our network outperforms state-of-the-art residual attention networks. Experimental results show that the performance of our proposed network is better than that of residual attention network.

II. ATTENTION MODULE

The visual attention mechanism is a brain signal processing mechanism that is unique to human vision [20] [21]. Human vision scans the global image quickly to identify the focus area, which is generally called the focus of attention, and then invests a large amount of attention resources in this area to obtain more details of the target and suppress other useless information. The attention mechanism in deep learning is essentially similar to the human selective visual attention mechanism. The core goal is to also select more information from the many information that is of vital importance to the current mission objectives [22]. In recent years, the attention model has been widely used in various types of deep learning tasks, such as natural language processing, image recognition and speech recognition [23] [24] [18]. It is a core technologies that deserves the most attention and insight. In this paper, the attention model used is based on the attention module of residual attention network.

To improve the classification accuracy, Wang fei et al. proposed a residual attention network [25]. Compared with ResNet-200 [26], the residual attention net achieves 0.6% top-1 accuracy improvement and has good robustness. Based on the above advantages, we adopt an attention module in a residual attention network. As shown in Figure 1, each attention module is divided into two branches: the mask branch and trunk branch. The trunk branch performs feature processing and can be adapted to any state-of-the-art network structure. For the mask, first, through a series of convolution and pooling, it gradually extracts high-level features and increases the receptive field of the model. Then the size of the feature map is enlarged to the same size as the original input by the same number of up samples. Thus, it maps the area of attention to each pixel.

Given input x, $T(x)$ denotes the trunk branch output, and the mask branch uses a bottom-up top-down structure to learn the same size of mask $M(x)$ that softly weights output features $T(x)$. The output mask is used as control gate for neurons of the trunk branch that are similar to the highway network. The output of attention module $H(x)$ is:

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

where i ranges over all spatial positions, c is the index of the channel($c \in \{1, \cdots, c\}$), $M(x)$ is the output of Soft Mask Branch and $T(x)$ is the output of the Trunk Branch. The entire structure can be trained end-to-end. Each pixel value in the attention map output by the mask branch is equivalent to the weight of each pixel value in the original feature map, which enhances meaningful features and suppresses meaningless information. Therefore, the weighted attention map is obtained by element-wise multiplication of the output of the mask branch and the output of the trunk branch. However, this weighted attention map cannot be directly input into the next layer because the activation function of the mask branch is sigmoid and the output value is in the range $(0, 1)$, so the author performs an element-wise operation on the weighted attention map and the feature map of the trunk branch.

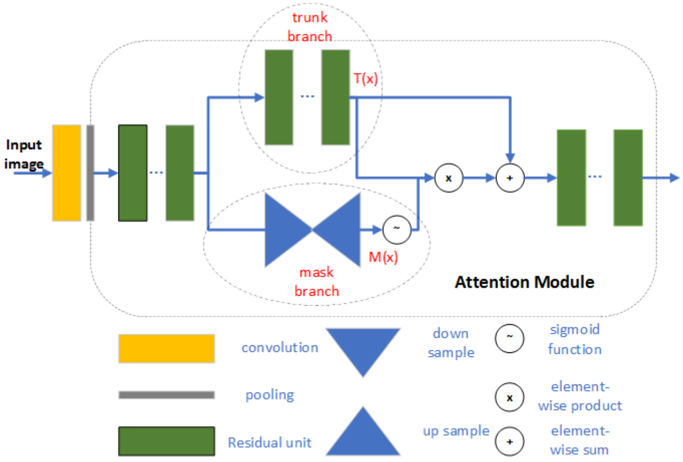

Fig. 1. Network structure of the attention module.

The attention module has the following advantages: 1) The attention module is similar to the residual learning mode, so a very deep model can be easily optimized and learned, and has very good performance. 2) The forward attention mechanism of bottom-up top-down. Other networks that use attention often need to add a branch to the original network to extract attention, and the author's model can extract the attention of the model in a forward process, thereby making the model training easier.

III. Method

In this paper, we construct a down-residual module and combine it with the attention module to construct an attention-based convolution network(ACNet). The network combines the advantages of the down-residual module with the advantages of the attention module to deliver outstanding performance and robustness.

A. Attention-based Network

We construct an attention-based convolution network based on attention. The overall network architecture and hyperparameter settings are shown in Figure 2. The network consists of three modules: the first module consists of seven convolutional layers and three residual blocks, and then two attention modules are connected. The former operation quickly collects the global information of the entire image and the latter operation extracts high-level distinguishable feature information.

From input x, convolution are firstly performed several times to increase the receptive field rapidly. After achieving the high-level features, they are used as input into the three residual layers. The residual layer can fully extract the useful advanced feature information and combine it with the input information of the upper layer. The useful feature information is fully used and missing useful feature information is avoided. Finally, the features are put into two attention layers.

B. Loss Function

Cross entropy describes the distance between two probability distributions. The smaller the cross entropy, the closer the two are. It is often used for the classication loss function of deep neural networks. In this paper, we use cross entropy as the loss function of the myocardial classification.

$$H(y, y') = -\sum_{i}^{n} y'_i log y_i, \tag{2}$$

where H is a representation of the loss value, y indicates the predicted probability distribution, y' denotes the true probability, and n suggests the number of samples. Furthermore, the cross entropy loss function is put to a frequent use in classification networks.

Learning rate strategy: We conducted several trials with different learning rates, and the results showed that the learning rate of *0.001* was the most appropriate. If the learning rate is too high, it will lead to over-fitting, and the network cannot learn the feature information correctly. If the learning rate is too low, the network fitting speed is very slow. Therefore we set the learning rate to *0.001*.

Experiment configurations: To ensure the consistency of the experiment, we use accuracy as the quantization metric, *100* epochs are trained for each experiment ($batch\_size$ = 3). All experiments are implemented in python2.7 by using Tensorflow and Keras framework. We train the networks on a NVIDIA Tesla M40 GPU and the model that performs the best on test data set are saved for further analysis.

IV. EXPERIMENT and Results

A. Dataset, Pre-processing and Evaluation Metrics

CMR T1-mapping images of 485 myocardial sections from the same hospital were both trained and tested. The ground truth of the training and the testing samples of images with the myocardial were manually annotated by an experienced cardiologist. The pixel size of each CMR image amounts 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 $1 \times 1 \times 1$, and each image was cut to a fixed image size of 128×128. Finally, the intensity range of the T1-mapping images was normalized to $[0 - 255]$.

To improve the classification performance of ACNet, three groups of experiments were constructed. In experiment *1*, we increased the dataset size by using *485* images with a size of 128 × 128(included 36 normal people and 449 patients), *388* training images, and *97* test images. In experiment *2*, because the ratio of the positive and negative samples of the above dataset was very unbalanced (the negative sample contained only *36* cases), the dataset may have caused the classification performance to decrease. To solve this problem, we used data enhancement methods such as image rotation and random increase of noise to increase the number of samples of normal people from *36* cases to *191* cases, a total of *640* datasets of size 128 × 128 [27]. A residual attention network, which is a broadly applied state-of-the-art network structure, was used as a baseline method. To conduct a fair comparison, we used most of the settings same as Residual Attention Network paper. We adopted the same weight initialization method as the previous study and trained residual attention network using nesterov SGD with a mini-batch size of *3*.

We used four indicators to assess the performance of the network, which includes accuracy, sensitivity and specificity. The accuracy is the average of the accuracy of in each group of experiments. In the medical field, 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:

$$sensitivity = \frac{A}{A + B}, \tag{3}$$

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

$$specificity = \frac{D}{C + D}, \tag{4}$$

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

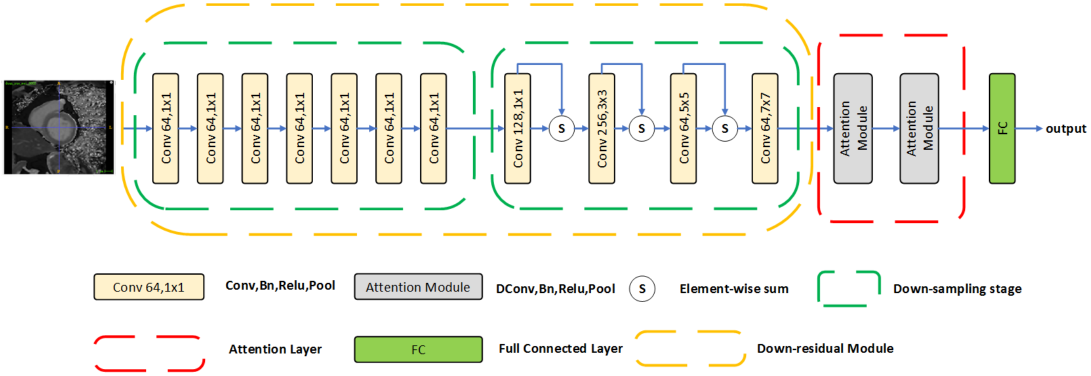

Fig. 2. The attention-based network structure proposed in this paper.

B. Results and Discussion

In this paper, we evaluated the performance of the residual attention network, VGG16, ResNet50, ResNet101 and our proposed network on the myocardial dataset. Two groups of experiments were constructed, and each group experiment was divided into two parts. In the first part, we used the myocardial dataset to perform multiple experiments on the residual attention network, VGG16, ResNet50 and ResNet101 to observe changes in the indicator. In the second part, we performed multiple experiments on the proposed attention-based network(ACNet) using the myocardial dataset to obtain the values of each indicator. Finally, the accuracy, sensitivity and specificity were compared to obtain the performance difference between the two methods.

For experiment *1*, as shown in Table I, in the case of the same amount of myocardial data, the classification accuracy of the residual attention network was only *0.4179*, the accuracy of VGG16 was only *0.8593*, the accuracy of resnet50 is only *0.6251*, and the accuracy of resnet101 is only *0.8347*, however, the accuracy of our proposed network can reached *0.9277*. Our approach is far superior to the other four comparison networks in terms of accuracy. Additionally, for the two indicators of sensitivity and specificity, our proposed network was also superior to the residual attention network, VGG16, ResNet50 and ResNet101. Unfortunately, the sensitivity of all methods in Table I was very low, whereas the specificity was close to *1*. This phenomenon was caused by the imbalance between the number of positive samples and the number of negative samples. Therefore, to solve this problem in experiment *2*, we used the data enhancement method to increase the number of negative samples, so that the difference between the positive and negative samples was reduced.

Figure 3 shows the attention features of two different samples in the residual attention network and ACNet. The myocardial instance mask highlighting the area surrounding of the myocardial. As can be observed from the figure, the myocardial mask highlighted the area around the myocardium and inhibited the middle area. The surrounding area of the myocardium is just the main feature area for assessing whether the heart muscle is normal. Therefore, we can know that the attention mechanism can highlighted useful feature information while suppressing the useless feature information, thereby improving the classification performance. Additionally, by comparing the feature maps of the two methods, we can clearly observe that ACNet's ability to highlight useful information and the ability to suppress useless information was stronger than that of the residual attention network. It essentially shows that ACNet outperformed the residual attention network.

TABLE I
EXPERIMENT 1 RESULTS FOR BOTH NETWORKS. IN THIS EXPERIMENT, A 5-FOLD CROSS-VALIDATION METHOD WAS USED. THE AVERAGE ACCURACY WAS THE AVERAGE OF FIVE EXPERIMENTAL PRECISIONS. THE SENSITIVITY AND SPECIFICITY WERE THE AVERAGE RESULTS OF FIVE EXPERIMENTS.

Network	Accuracy	sensitivity	specificity
Residual attention network	0.4179	0.1067	0.9571
VGG16	0.8593	0.3642	0.9142
ResNet50	0.6251	0.1753	0.8327
ResNet101	0.8347	0.3162	0.9059
Ours	0.9277	0.5305	0.9901

For experiment *2*, as shown in Table II, we used the data enhancement method to increase the number of negative samples. This method can reduce the difference in the number of between positive and negative samples, the number of positive samples and the number of negative samples remained relatively balanced. From Table II, we can observe that ACNet's three indicators were much higher than those of residual attention network, VGG16, ResNet50 and ResNet101. This result once again demonstrates that the performance of our proposed method is the most outstanding.

Figure 4 shows the comparison of the accuracy results of the residual attention network, VGG16, ResNet50, ResNet101 and ours ACNet. From the figure, we can clearly observe that, in the case of the same DCM dataset, the ACNet classification accuracy was always much higher than that of remaining four networks. From Figure 5, we can observe that the specificity of ACNet was much larger than that of remaining four networks, indicating that ACNet's ability to predict correct negative

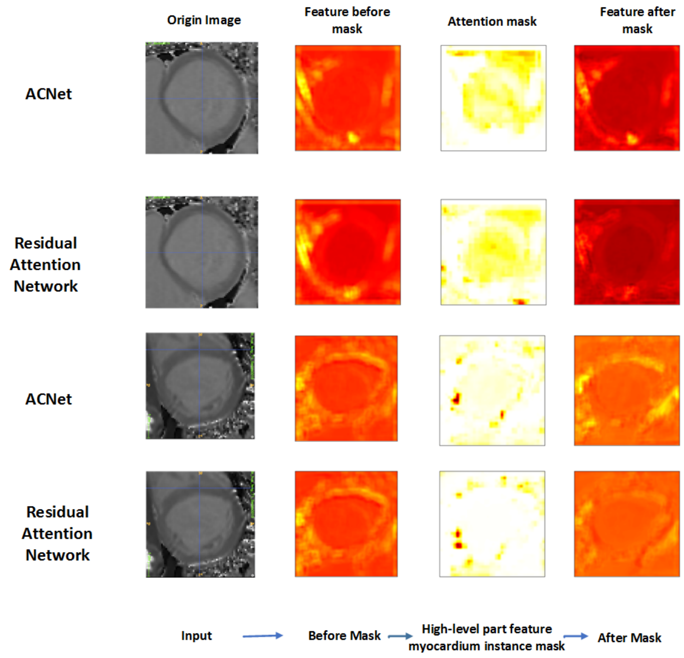
Origin Image
Feature before mask
Attention mask
Feature after mask
ACNet
Residual Attention Network
ACNet
Residual Attention Network
Input
Before Mask
High-level part feature myocardium instance mask
After Mask

Fig. 3. Example images illustrating that different features have different corresponding attention masks in attention module. The myocardium instance mask highlights high-level myocardium surrounding part features.

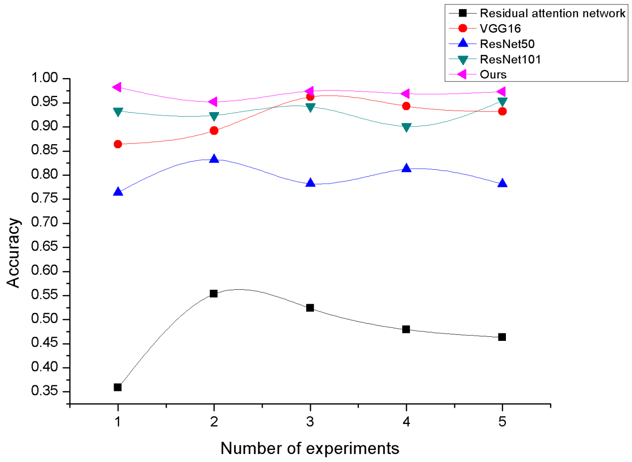

Fig. 4. Distribution of accuracy results in experiment 2.

samples was significantly higher than that of remaining four networks. The difference in sensitivity results between the residual attention network, VGG16, ResNet50, ResNet101 and ACNet is shown in Figure 6. From the figure we can observe that the sensitivity result of the residual attention network, VGG16, ResNet50 and ResNet101 was much lower than that of ACNet. This demonstrates that the ability of remaining four comparison networks to predict correct positive samples was far lower than that of ACNet.

In general, from a comprehensive analysis of the results of the above methods, we can clearly observe that the advantages

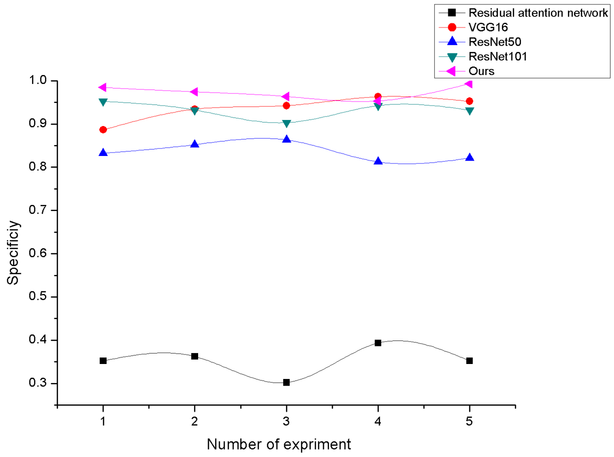

Fig. 5. Distribution of specificity results in experiment 2.

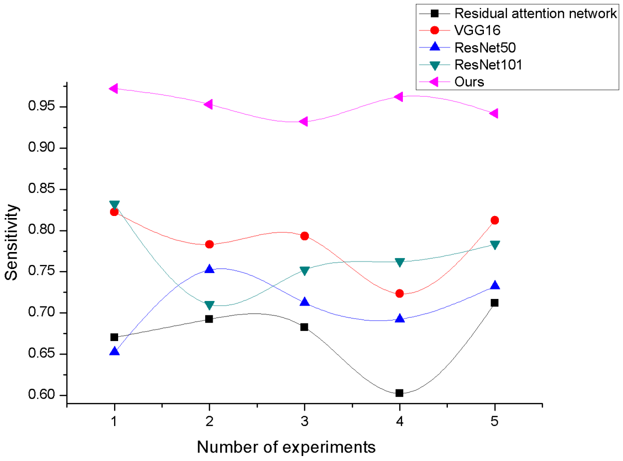

Fig. 6. Distribution of sensitivity results in experiment 2.

of ACNet proposed in this paper were very prominent. Our method is much higher than the four comparison methods in terms of accuracy, sensitivity and specificity. From Figure 4 6 5, we can see that the three indicators of our method are very stable. It can be seen that our method is very robust. In the medical image classification task, the performance of ACNet was far stronger than that of the residual attention network, VGG16, ResNet50 and ResNet101. From the many experimental results, we can observe that the experimental results of ACNet were very prominent and stable, the robustness of ACNet was also strong.

To summarize, in the two groups of experiments, the four indicators of ACNet were much larger than those of the remaining four comparison networks. Thus, we can clearly observe that ACNet outperformed the residual attention network, VGG16, ResNet50 and ResNet101. Furthermore, the three indicators of experiment *2* were much higher than those of experiment *1*, in particular, the results of ACNet in experiment

TABLE II
EXPERIMENT 2 RESULTS OF BOTH NETWORKS. IN THIS EXPERIMENT, THE UP-SAMPLING METHOD WAS USED TO INCREASE THE NUMBER OF NEGATIVE SAMPLES.

Network	Accuracy	Sensitivity	Specificity
Residual attention network	0.4718	0.6890	0.3510
VGG16	0.9452	0.7891	0.9405
ResNet50	0.7859	0.7132	0.8327
ResNet101	0.9274	0.7672	0.9361
Ours	0.97152	0.9582	0.9761

2 were significantly higher than the results of experiment *1*. Therefore, we found that after using the data enhance method, performance was greatly improved. Similarly, in the field of medical imaging, there are often problems such as small datasets, and unbalanced proportions of positive and negative samples. We can use the method of data enhance to appropriately increase the number of samples, so that the number of positive and negative samples is basically the same. This can greatly improve the classification performance of the network.

V. CONCLUSION

In this paper, we proposed an attention-based convolution network. It combines attention module and down-residual module designed that we designed. The down-residual module can acquire high-level features without losing the original features, and the attention module highlights useful features and suppresses useless features. To prove the effectiveness of our framework, we compared it with residual attention network, VGG16, ResNet50 and ResNet101. We found that the proposed method was superior to remaining four comparison networks in terms of three indicators. The experimental results show that this method is obviously superior to the most advanced classification methods, with the highest accuracy, sensitivity and specificity. Additionally, the framework that we used is generally applicable to the classification tasks of other medical images or natural images. We plan to conduct further research in the future.

REFERENCES

[1] C. Tian, *A Computer Vision-Based Classification Method for Pearl Quality Assessment*, 2009.

[2] N. Doulamis and A. Doulamis, "Semi-supervised deep learning for object tracking and classification," in *IEEE International Conference on Image Processing*, 2015, pp. 848–852.

[3] S. Pang, J. J. D. Coz, Z. Yu, O. Luaces, and J. Dłez, *Combining Deep Learning and Preference Learning for Object Tracking*. Springer International Publishing, 2016.

[4] F. L. C. D. Santos, M. Paci, L. Nanni, S. Brahnam, and J. Hyttinen, "Computer vision for virus image classification," *Biosystems Engineering*, vol. 138, pp. 11–22, 2015.

[5] T. M. Lehmann, H. Schubert, D. Keysers, M. Kohnen, and B. B. Wein, "The irma code for unique classification of medical images," in *Medical Imaging*, 2003.

[6] I. Buciu and A. Gacsadi, "Gabor wavelet based features for medical image analysis and classification," in *International Symposium on Applied Sciences in Biomedical and Communication Technologies*, 2009, pp. 1–4.

[7] C. Grefkes, S. B. Eickhoff, D. A. Nowak, M. Dafotakis, and G. R. Fink, "Dynamic intra- and interhemispheric interactions during unilateral and bilateral hand movements assessed with fmri and dcm," *Neuroimage*, vol. 41, no. 4, pp. 1382–1394, 2008.

[8] N. K. Lakdawala, L. Dellefave, C. S. Redwood, E. Sparks, A. L. Cirino, S. Depalma, S. D. Colan, B. Funke, R. S. Zimmerman, and P. Robinson, "Familial dilated cardiomyopathy caused by an alpha-tropomyosin mutation : The distinctive natural history of sarcomeric dilated cardiomyopathy," *Journal of the American College of Cardiology*, vol. 55, no. 4, pp. 320–329, 2010.

[9] H. D. Theiss, D. Robert, M. G. Engelmann, B. Andreas, S. Klaus, N. Michael, R. Bruno, S. Gerhard, and F. Wolfgang-M, "Circulation of cd34+ progenitor cell populations in patients with idiopathic dilated and ischaemic cardiomyopathy (dcm and icm)," *European Heart Journal*, vol. 28, no. 10, p. 1258, 2007.

[10] X. Li, J. C. Lv, and Y. Zhang, "Manifold alignment based on sparse local structures of more corresponding pairs," in *Twenty-Third International Joint Conference on Artificial Intelligence*, 2013.

[11] X. Li, J. Lv, and Z. Yi, "An efficient representation-based method for boundary point and outlier detection," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 1, pp. 51–62, 2018.

[12] ——, "Outlier detection using structural scores in a high-dimensional space," *IEEE transactions on cybernetics*, 2018.

[13] X. Li, J. Lv, X. Wu, and X. Yu, "A semi-supervised manifold alignment algorithm and an evaluation method based on local structure preservation," *Neurocomputing*, vol. 224, pp. 195–203, 2017.

[14] Y. Jiang, Z. Li, L. Zhang, and P. Sun, *An Improved SVM Classifier for Medical Image Classification*. Springer Berlin Heidelberg, 2007.

[15] B. Li and Q. H. Meng, "Tumor ce image classification using svm-based feature selection," in *Ieee/rsj International Conference on Intelligent Robots and Systems*, 2010, pp. 1322–1327.

[16] G. M. Foody and A. Mathur, "Toward intelligent training of supervised image classifications: directing training data acquisition for svm classification," *Remote Sensing of Environment*, vol. 93, no. 1, pp. 107–117, 2004.

[17] D. S. Deshpande, A. M. Rajurkar, and R. M. Manthalkar, "Medical image analysis an attempt for mammogram classification using texture based association rule mining," in *Computer Vision, Pattern Recognition, Image Processing and Graphics*, 2014, pp. 1–5.

[18] H. I. Kim and R. H. Park, "Residual lstm attention network for object tracking," *IEEE Signal Processing Letters*, vol. PP, no. 99, pp. 1–1, 2018.

[19] Y. Gao, Y. Chen, J. Wang, and H. Lu, "Reading scene text with attention convolutional sequence modeling," 2017.

[20] Y. Sun and R. Fisher, "Object-based visual attention for computer vision," *Artificial Intelligence*, vol. 146, no. 1, pp. 77–123, 2003.

[21] F. Wang and D. M. J. Tax, "Survey on the attention based rnn model and its applications in computer vision," 2016.

[22] F. Miau, C. S. Papageorgiou, and L. Itti, "Neuromorphic algorithms for computer vision and attention," in *International Symposium on Optical Science and Technology*, 2001.

[23] M. M. Mahsuli and R. Safabakhsh, "English to persian transliteration using attention-based approach in deep learning," in *Electrical Engineering*, 2017, pp. 174–178.

[24] H. Wang, S. Tang, Y. Zhang, T. Mei, Y. Zhuang, and F. Wu, "Learning deep contextual attention network for narrative photo stream captioning," in *on Thematic Workshops of ACM Multimedia*, 2017, pp. 271–279.

[25] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Computer Vision and Pattern Recognition*, 2017, pp. 6450–6458.

[26] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," 2016.

[27] D. Balasubramanian, M. C. Krishna, and R. Murugesan, "Convolution-based interpolation kernels for reconstruction of high resolution emr images from low sampled k-space data," in *International Conference on Conference on Computational Intelligence and Multimedia Applications*, 2007, pp. 308–313.