2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

A Self-supervised Learning Based Framework for Automatic Heart Failure Classification on Cine Cardiac Magnetic Resonance Image

Hai Zhong, Jiaqi Wu, Wangyuan Zhao, Xiaowei Xu, Runping Hou, Lu Zhao, Ziheng Deng, Min Zhang, and Jun Zhao* *Member, IEEE*

Abstract— **Heart failure (HF) is a serious syndrome, with high rates of mortality. Accurate classification of HF according to the left ventricular ejection faction (EF) plays an important role in the clinical treatment. Compared to echocardiography, cine cardiac magnetic resonance images (Cine-CMR) can estimate more accurate EF, whereas rare studies focus on the application of Cine-CMR. In this paper, a self-supervised learning framework for HF classification called SSLHF was proposed to automatically classify the HF patients into HF patients with preserved EF and HF patients with reduced EF based on Cine-CMR. In order to enable the classification network better learn the spatial and temporal information contained in the Cine-CMR, the SSLHF consists of two stages: self-supervised image restoration and HF classification. In the first stage, an image restoration proxy task was designed to help a U-Net like network mine the HF information in the spatial and temporal dimensions. In the second stage, a HF classification network whose weights were initialized by the encoder part of the U-Net like network was trained to complete the HF classification. Benefitting from the proxy task, the SSLHF achieved an AUC of 0.8505 and an ACC of 0.8208 in the 5-fold cross-validation.**

I. Introduction

Heart failure (HF) is a serious condition with a high prevalence rate of about 2% in developed countries and more than 7% over 75 years of age in these countries [1]. Clinically, left ventricular ejection fraction (EF) less than 50% is considered to be heart failure with reduced ejection fraction (HFrEF), and EF more than 50% is considered to be heart failure with preserved ejection fraction (HFpEF) [2]. They differ greatly in clinical treatment due to their different biological characteristics [2]. Thus, the classification of HFrEF and HFpEF are quite important for clinical decision making.

In clinical, EF evaluation generally requires manually delineating the left ventricle area during the end-systolic and end-diastolic, and then calculating the volume differences. However, manual delineation of the left ventricle is time-consuming, laborious and subjective, thus we aim to construct an automatic classification model to reduce the delineation burden of clinicians.

In recent years, machine learning and artificial intelligence algorithms have shown great potential in the medical field. At present, there are many kinds of researches on HF based on machine learning. For example, in the automatic diagnosis of HF, studies [3-5] have proposed different machine learning algorithms to diagnose HF based on electrocardiograph (ECG) and achieved satisfactory results. In the automatic classification of HF, Delaram et al. [6] used echocardiography to train a dual-channel deep neural network for EF estimation. However, because of the poor quality of echocardiography and the unclear boundary of the cardiac region, using echocardiography to estimate EF may not be accurate enough.

Compared with echocardiography, cine cardiac magnetic resonance images (Cine-CMR) can not only clearly display the boundary of the heart because of its high resolution, but also show the heart beating during the entire cardiac cycle due to its 4D property, all of which is beneficial to the accurate evaluation of EF. Therefore, this paper used Cine-CMR to construct an automatic classification model.

At present, constructing reliable computer-aided diagnosis systems depends on sufficient labeled data, whereas it is difficult to obtain a mass of labeled medical data. This research is faced with the similar problem. Self-supervised learning provides an opportunity to train a reliable model with a small dataset. Therefore, in this paper, we intended to use the self-supervised learning method to ameliorate the problem of data deficiencies. For the self-supervised learning methods in medical images, there has been some existing studies using the spatial information to construct proxy task. For example, Zhou et al. [7] and Chan et al. [8] have proposed similar image restoration proxy tasks to enable the network learn the spatial features of the medical images in advance. Whereas, in order to learn the information of EF from Cine-CMR, the network not only needs to learn the spatial information but also needs to learn the temporal information. However, few studies have focused on using self-supervised learning to help mine the temporal information of medical images. Some temporal self-supervised methods have been proposed in the nature video processing field. For instance, studies [9-10] have proposed different self-supervised methods (correct order judgement and jigsaw puzzles) to help network exploit the temporal information. These methods are effective when the motion range in the video is large enough, however, the heartbeats motion is in a small range. Consequently, how to design a self-supervised learning proxy task which can combine the spatial information and the temporal information (in a small motion range) in medical images remains challenging.

In this paper, a self-supervised learning based framework for HF classification called SSLHF (Fig.1) was proposed to complete HF classification based on a small amount of data.

Research supported by Shanghai Hospital Development Center Clinical Science and Technology Innovation project (SHDC12019X22).

Hai Zhong, Wangyuan Zhao, Xiaowei Xu, Runping Hou, Lu Zhao, Ziheng Deng, and Jun Zhao are with School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China.

Jiaqi Wu and Min Zhang is with Department of Cardiology, Shanghai Chest Hospital, Shanghai, 200000, China.

Corresponding to: Jun Zhao (junzhao@sjtu.edu.cn).

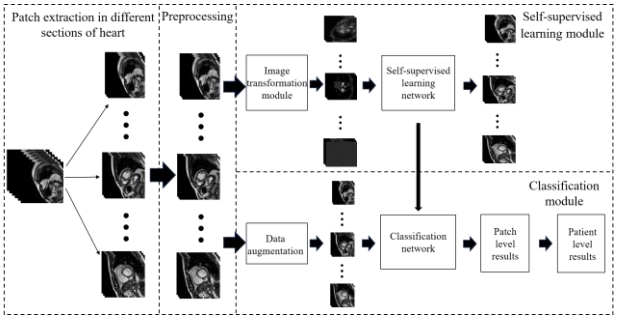


Figure 1. The structure of proposed SSLHF.

Specially, the self-supervised learning module can help the classification network exploit the spatial and temporal information. Firstly, we extracted patches according to different sections of the heart. Secondly, Z-score normalization was used to normalize the intensity of the patches. Then, the patches were used as supervised labels to train an image restoration network. Finally, we transferred the structures and weights of the encoder from the self-supervised network to the classification network and used the HF category labels to fine-tune the classification network. Details of the steps will be illustrated in Section III.

II. MATERIALS

In our study, Cine-CMR from 184 patients diagnosed with HF were collected from Shanghai Chest Hospital, including 63 patients diagnosed with HFrEF and 121 patients diagnosed with HFpEF. The number of sections of heart is different among Cine-CMR of different patients and each section has 40 frames to reflect the motion of the heart. In our experiment, we extracted the patches from Cine-CMR according to different sections of heart. The final samples used in this paper include 480 HFrEF patches and 895 HFpEF patches. These patches were randomly divided into 5 folds (to prevent data leakage, patches from the same patient would only be divided into the same fold), 3 folds for training, 1 fold for validating and 1 fold for testing. Then 5-fold cross-validation was used to let each fold as the test set for more reliable results. Ethics approval was obtained from the Shanghai Chest Hospital, School of Medicine, Shanghai Jiao Tong University.

III. METHODS

A. Extracting Patches and Preprocessing

In order to unify the input size of neural network and increase the number of training samples, we extracted patches with size of 288×288×40 (288 is the size of x and y, 40 is the number of frames) from Cine-CMR. Patches which do not include the left ventricular region were eliminated with the guidance of a physician. The pixel values of patches were truncated by a threshold of 2000 and then Z-score normalization was employed to normalize all the patches into a similar intensity range.

B. Image Restoration Proxy Task

For the purpose of ameliorating the problem of data deficiencies and helping the classification network better mine the spatial and temporal features of Cine-CMR, we designed an image restoration proxy task. In this stage, a U-Net like network (Fig.2 Part A and Part B) was proposed for image restoration. The encoder of the network contained ten 3D convolution layers and four 3D max-pooling layers. The decoder of the network contained ten 3D convolution layers and four 3D transpose convolution layers.

The key to the image restoration task is how to set image transformation. In order to enable the classification network learn the 2D spatial information and the temporal information in advance, three methods of image transformation were proposed as follows:

- Local pixel shuffling [7]: For each frame, we sampled some small enough windows (smaller than the receptive field) randomly from the frame and then shuffled the pixels inside each window sequentially. The local pixel shuffling can be formulated as follows:

$$\tilde{W}=\bar{P}\times W\times P \tag{1}$$

  where $\tilde{W}$ and W are the transformed window and sampled window respectively. $\bar{P}$ and P denote permutation metrics with the size of $n\times n$ and $m\times m$. This method aims to help the network learn the texture information from restoring the correct pixels order.

- Patch covering: Given an original frame, we selected some candidate patches randomly from the frame. Then we used random values to replace the pixel values of the candidate patches. In order to control the difficulty of restoration task and guarantee the effect, the area of selected patches is bigger than one thirty-sixth and smaller than one ninth of original frame. Besides, no more than 5 patches are selected from the same frame. This method is designed to make the network learn the 2D spatial context information from restoring the patches.

- Frame covering: Given a training sample, we selected five groups of consecutive frames randomly. Each group contains no more than three frames. Then we used random values to replace the pixel values in the selected frames. The purpose of frame covering is to help the network learn the temporal information from restoring the frame. Moreover, the restoration task is a pixel-to-pixel task, thus this transformation method can capture the small motion of the heart.

The training samples with one or more transformations were used as the inputs of the image restoration network and the original training samples were used as labels to train the network. After the image restoration network has been trained, the weights can be transferred into the classification network in the next stage.

C. HF Classification

In this stage, a classification network (Fig.2 Part C) was proposed for HF classification. The backbone of the classification network is the same as the encoder architecture of the U-Net like network, and a max-pooling layer and three fully connection layers were appended to this backbone for aggregating features and outputting the final predictions.

After the U-Net like network has been trained, the pre-trained parameters of the encoder were used to initialize the classification network and the HF classification labels were applied as supervised information to fine-tune the classification network.

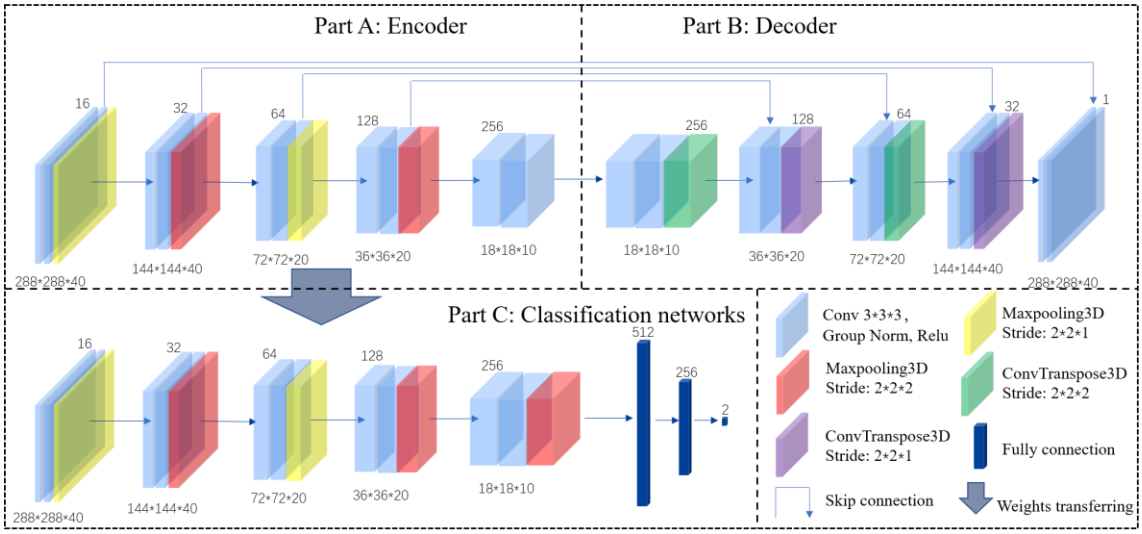


Figure 2. The architectures of the networks employed in this paper. Part A is the architectures of the encoder in self-supervised network. Part B is the architectures of the decoder in self-supervised network. Part C is the architectures of classification network.

Owing to the data imbalance in this dataset, weighted binary cross-entropy loss was utilized to train the network. The formula of the weighted binary cross-entropy loss was shown below:

$$L_{WBCE} = -\frac{1}{N}\sum_{i=1}^{N} w_i \times [y_i \times \ln \hat{y}_i + (1-y_i) \times \ln(1-\hat{y}_i)] \quad (2)$$

where w_i is the weight of different classes. y_i and $\hat{y}_i$ are the label and predicted value respectively and N is the batch size.

D. Aggregating Patches

The patch-level predictions can be obtained after the HF classification has been trained. In clinical, physicians pay more attention to the patient-level prediction results, thus it is necessary to aggregate patch-level predictions into patient-level predictions. In our study, we used the average pooling method to get patient-level predictions, the formula was shown below:

$$P_i = \frac{1}{n_i}\sum_{j=1}^{n_i} p_{ij} \quad (3)$$

Where p_{ij} and P_i are the patch-level prediction and patient-level prediction respectively, and n_i is the number of patches.

IV. Results And Discussion

A. Implementation Details

In the first stage, Adam optimizer was employed to minimize the mean-squared loss function. The learning rate was initialized to 0.001 and decayed by 0.5 every twenty epochs. Data augmentation including horizontal flips, vertical flips, time flips and rotations around the time axis was employed to reduce overfitting.

In the second stage, we used the weights of the encoder in the first stage to initialize the convolution layers and used the Kaiming method [11] to initialize the fully connection layers. The learning rate was initialized to 0.00001 and decayed by 0.5 every ten epochs to fine-tune the classification network. The optimizer and data augmentation methods were the same as those in the first stage.

B. Image Restoration Proxy Task

Fig. 3 shows the visualization results of the image restoration task, where the first row represents the restoration result of the spatial transformations (local pixel shuffling and patch covering) and the second row represents the restoration result of the temporal transformation (frame covering). According to the first row, obviously although the spatial information of the transformed image is missing or disordered (Fig.3 (b)), the self-supervision network can also get adequate restoration result (Fig.3 (c)), which indicates that the image restoration network has learned the texture information and the 2D spatial context information. In addition, according to the result of frame covering, we can see that even though no information exists in the original frame, the restoration result (Fig.3 (f)) is also acceptable. This restoration can only rely on previous frames or posterior frames, which means the image restoration network has learned the temporal information.

C. HF Classification

The classification results (the average results of 5-fold cross validation) are tabulated in Table I. As we can see from Table I, the AUCs and ACCs of patch-level and patient-level are both satisfactory. Due to the aggregation method reducing

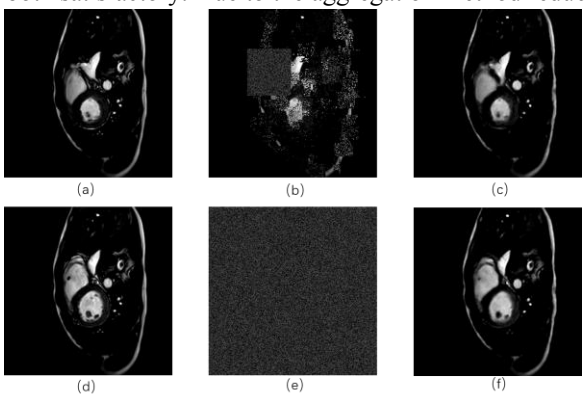


Figure 3. The results of image restoration network. (a) and (d) are the ground truth (image before transformation). (b) is a single frame transformed with local pixel shuffling and patch covering. The (e) is a single frame transformed with frame covering. (c) and (f) are the image restoration results of (b) and (e) respectively.

TABLE I. QUANTITATIVE ASSESSMENT OF CLASSIFICATION METHOD

Patch-level AUC	Patch-level ACC	Patient-level AUC	Patient-level ACC
0.8289	0.7682	0.8505	0.8208

the influence of outlier, the classification performance in patient-level is better than that in patch-level.

D. Ablation Studies

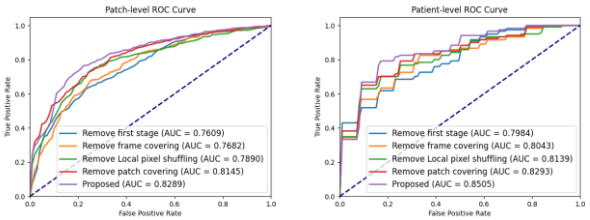
In order to explore the effects of self-supervised learning task on the classification task and evaluate the impact of different transformation methods, the ablation studies were conducted. We designed four ablation studies: A. remove the first stage (used the Kaiming method [11] to initialize the classification network), B. remove frame covering, C. remove local pixel shuffling and D. remove patch covering. As we can see from Fig. 4 and Table II, self-supervised learning module can greatly improve HF classification performance and each transformation method is effective. Notably, because of the significance of the temporal information in HF classification, removing frame covering has the biggest impact on performance.

E. Comparative Experiment

Because few studies have focused on the automatic classification of HF based on Cine-CMR, we compared our SSLHF with classic classification networks: 3D Resnet50 and 3D Densenet121. Due to the memory limitation, the batch size can only be set to 4. Consequently, the batch normalization layers in the network structure were changed to the group normalization layers in our study. According to Table III, our proposed methods show great superiority in HF classification, indicating that comparing to classic classification networks, our framework is effective in HF classification task.

F. Limitations

The proposed methods only take into account the 2D spatial information and the temporal information, but ignore the 3D spatial information among different heart sections.



(a) (b)

Figure 4. The ROC curves for HF classification on patch-level and patient-level. (a) The 5-fold cross-validation mean ROC curves for different ablation studies on patch-level. (b) The 5-fold cross-validation mean ROC curves for different ablation studies on patient-level.

TABLE II. QUANTITATIVE ASSESSMENT OF ABLATION STUDIES

	Patch-level AUC	Patch-level ACC	Patient-level AUC	Patient-level ACC
Remove first stage	0.7609	0.7178	0.7984	0.7496
Remove frame covering	0.7682	0.7328	0.8043	0.7774
Remove local shuffling	0.7890	0.7518	0.8139	0.7826
Remove patch covering	0.8145	0.7650	0.8293	0.8204
SSLHF	**0.8289**	**0.7682**	**0.8505**	**0.8208**

TABLE III. QUANTITATIVE ASSESSMENT OF DIFFERENT CLASSIFICATION METHODS

	Patch-level AUC	Patch-level ACC	Patient-level AUC	Patient-level ACC
3D Resnet50	0.7450	0.7100	0.7678	0.7614
3D Densenet121	0.7145	0.7052	0.7822	0.7614
SSLHF	**0.8289**	**0.7682**	**0.8505**	**0.8208**

However, the 3D spatial information is also important for HF classification. Because of the limited amount of data, it is difficult to design an algorithm to combine the 3D spatial information and the temporal information directly. Therefore, how to combine the 3D spatial and temporal information in a small dataset is the direction of future research.

V. CONCLUSION

In this paper, SSLHF was proposed to automatically distinguish HFpEF from HFrEF. In order to ameliorate the problem of data deficiencies, a self-supervised learning module was designed in this framework. Combined with the characteristics of Cine-CMR, we proposed three different image transformations in this module to help the network learn the texture information, the 2D spatial context information and the temporal information (in a small motion range) in advance. This proposed SSLHF achieved an AUC of 0.8505 and an ACC of 0.8208 in patient-level, showing its potential in assisting physicians in personalized treatment plans.

REFERENCES

[1] A. Mosterd and A. W. Hoes, "Clinical epidemiology of heart failure," *Heart*, vol. 93, no. 9, pp. 1137-46, Sep 2007.

[2] J. J. McMurray et al., "ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure 2012: The Task Force for the Diagnosis and Treatment of Acute and Chronic Heart Failure 2012 of the European Society of Cardiology. Developed in collaboration with the Heart Failure Association (HFA) of the ESC," *Eur Heart J*, vol. 33, no. 14, pp. 1787-847, Jul 2012.

[3] Z. Masetic and A. Subasi, "Congestive heart failure detection using random forest classifier," *Comput Methods Programs Biomed*, vol. 130, pp. 54-64, Jul 2016.

[4] U. R. Acharya et al., "Deep convolutional neural network for the automated diagnosis of congestive heart failure using ECG signals," *Applied Intelligence*, vol. 49, no. 1, pp. 16-27, 2018.

[5] J. M. Kwon et al., "Development and Validation of Deep-Learning Algorithm for Electrocardiography-Based Heart Failure Identification," *Korean Circ J*, vol. 49, no. 7, pp. 629-639, Jul 2019.

[6] D. Behnami et al., "Automatic cine-based detection of patients at high risk of heart failure with reduced ejection fraction in echocardiograms," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 8, no. 5, pp. 502-508, 2019.

[7] Z. Zhou, V. Sodha, J. Pang, M. B. Gotway, and J. Liang, "Models Genesis," *Med Image Anal*, vol. 67, p. 101840, Jan 2021.

[8] L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara, and D. Rueckert, "Self-supervised learning for medical image analysis using image context restoration," *Med Image Anal*, vol. 58, p. 101539, Dec 2019.

[9] I. Misra, C. L. Zitnick, and M. Hebert, "Shuffle and learn: unsupervised learning using temporal order verification," in *European Conference on Computer Vision*, pages 527–544, springer, 2016.

[10] U. Ahsan, R. Madhok, and I. Essa, "Video Jigsaw: Unsupervised Learning of Spatiotemporal Context for Video Action Recognition," presented at *the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1026-1034.