Regional Cardiac Motion Scoring with Multi-scale Motion-based Spatial Attention

Wufeng Xue, Zejian Chen, Tianfu Wang, Shuo Li* and Dong Ni*

***Abstract*— Regional cardiac motion scoring aims to classify the motion status of each myocardium segment into one of the four categories (normal, hypokinetic, akinetic, and dyskinetic) from multiple short-axis MR sequences. It is essential for prognosis and early diagnosis for various cardiac diseases. However, the complex motion procedure of the myocardium and the invisible pattern differences pose great challenges, leading to low performance for automatic methods. Most existing works mitigate the task by differentiating the normal motion patterns from the abnormal ones, without fine-grained motion scoring.**

We propose an effective method for the task of cardiac motion scoring by connecting a bottom-up and another top-down branch with a novel motion-based spatial attention module in multi-scale space. Specifically, we use the convolution blocks for low-level feature extraction that acts as a bottom-up mechanism, and the task of optical flow for explicit motion extraction that acts as a top-down mechanism for high-level allocation of spatial attention. To this end, a newly designed *Multi-scale Motion-based Spatial Attention* (MMSA) mechanism is used as the pivot connecting the bottom-up part and the top-down part, and adaptively weight the low-level features according to the motion information.

Experimental results on a newly constructed dataset of 1440 myocardium segments from 90 subjects demonstrate that the proposed MMSA can accurately analyze the regional myocardium motion, with accuracies of 79.3% for 4-way motion scoring, 89.0% for abnormality detection, and correlation of 0.943 for estimation of motion score index. This work has great potential for practical assessment of cardiac motion function.

***Index Terms*— cardiac motion scoring, optical flow, attention mechanism, cardiac MR**

I. Introduction

MYOCARDIAL wall motion scoring plays a critical role in the clinical diagnosis and prognosis of various heart diseases for its sensitivity to early-stage functional alterations, including coronary artery disease, congestive heart disease, stress-induced cardiomyopathy, myocarditis, stroke, *etc*. [1]. However, accurate and robust motion analysis from cardiac imaging is a very challenging task due to the complex deformation procedure of cardiac movement with region wall thickening, circumferential shortening, and longitudinal ventricular shortening. This makes regional myocardial motion heterogeneous and location dependent [2].

In this work, we propose to conduct cardiac regional motion scoring from short-axis CMR cine sequences. Cardiac motion scoring aims to accurately grade the motion status of each myocardium segment, and is an essential part of patients' treatment decisions and rehabilitation evaluation. In routine clinical practice, motion scoring is often conducted by labor-intensive visual inspection of the dynamic cardiac sequence for each segment of left-ventricle myocardium according to AHA 17-segments model, following the 4-way scoring system of [1]: 1) normal, 2) hypokinetic, 3) akinetic, and 4) dyskinetic. The patterns of different segments are complex and the subtle differences between these patterns are difficult to identify. As a result, the motion score obtained by labor-intensive visual inspection is usually characterized by large inter-rater variability and low reproducibility [3].

With the development of machine learning in medical images analysis, automatic methods can be used to help analysis of the myocardium motion. Most of the previous efforts on cardiac motion analysis focused on binary motion abnormality detection [4]–[16], i.e, differentiating normal motion from abnormal motion, which is much easier than the four-way motion scoring task. Nevertheless, these methods suffer from several limitations (see *Related work in* II-A) that they cannot be directly applied or extended to the task of motion scoring.

As our preliminary attempt for this task, a deep convolution neural network (CNN), which we denoted as Cardiac-MOS, was proposed in [17], and achieved state-of-the-art performance. It leverages the powerful representation of CNN to capture the discriminate features from each frame, and models the long-range dependence of the distant non-local responses to identify the subtle differences between myocardium segments. However, the method used only one convolution neural network for the weakly labeled sequence-level classification and may be insufficient to obtain discriminative features with spatial and temporal characteristics. This inspires us to explicitly leverage motion information for the task.

In this work, we propose to combine a bottom-up mechanism with a top-down mechanism for the task of motion

Wufeng Xue, Zejian Chen, Tianfu Wang and Dong Ni are with the National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, Medical Ultrasound Image Computing (MUSIC) Laboratory, School of Biomedical Engineering, and Marshall Laboratory of Biomedical Engineering, Health Science Center, Shenzhen University, Shenzhen, China.

Shuo Li is with the Department of Medical Imaging, Western University, London, ON N6A 3K7, Canada.

The work is partially supported by the Natural Science Foundation of China (62171290, 61801296), the Shenzhen Basic Research (JCYJ20190808115419619), Shenzhen-Hong Kong Joint Research Program (No. SGDX20201103095613036), Medical Scientific Research Foundation of Guangdong Province (No. A2021370), and the Shenzhen College Stable Support Plan (20200812162245001).

Corresponding author: Dong Ni (nidong@szu.edu.cn), Shuo Li (slishuo@gmail.com)

scoring. Specifically, we 1) firstly use the convolution blocks for low-level feature extraction, which acts as a bottom-up mechanism, and 2) then use the task of unsupervised optical flow for explicit motion information extraction, which acts as a top-down mechanism for high-level allocation of spatial attention. To this end, a newly designed *Multi-scale Motion-based Spatial Attention* (MMSA) module is used as the pivot to adaptively weigh the low-level features according to the motion information in multi-scale space, and therefore connecting the bottom-up part and the top-down part. Besides, in addition to the non-local module (as used in [17]) for long-range information aggregation, a temporal squeeze-and excitation module is also introduced to enhance feature effectiveness with temporal weighting.

Compared to our preliminary work [17], the proposed method has advantages of 1) explicitly extracting motion information of myocardium by an unsupervised optical flow estimation task, 2) regularizing the parameter space with the top-down motion extraction task, and learning more discriminant motion-aware features with the newly proposed MMSA module, and 3) enhancing the bottom-up convolution features by non-local dependency modeling and temporal squeeze and excitation.

The contribution of this work is threefold:

- We propose an effective deep learning-based cardiac motion scoring method, which is capable of predicting the motion status of 16 myocardium segments from slices of short-axis cardiac cine MR sequences, and has great potential for assessment of cardiac dynamic function.
- We design a mechanism of multi-scale motion-guided discriminative representation learning, which connects a bottom-up convolutional representation part with a top-down motion extraction part through a newly-designed motion-based spatial attention module in multi-scale.
- We construct the first cardiac cine MR database for regional motion scoring. The dataset provides a platform for myocardium motion-related work, such as regional motion scoring, motion score index estimation, and abnormality detection. It will attract wide interest for advanced cardiac motion scoring tasks.

The rest of this paper is organized as follows: In Section II, we describe existing work for myocardium motion analysis and techniques for motion extraction of sequential data. Then, in Section III we describe our method in detail, including data preprocessing, the network architecture, and different modules used in the method. In Section IV, we introduce the dataset and the experimental settings. In Section V, we validate the effectiveness of each module, show the results of our method for myocardium motion analysis, and make comparisons with state-of-the-art methods. Section VI discusses the key factors for the task of motion scoring and the potential direction to improve the work on cardiac motion analysis. Finally, we give our conclusion in Section VII.

II. Related Work

A. Myocardium motion analysis

Many efforts have been devoted to automatic myocardium motion analysis by leveraging the discriminative power of supervised machine learning methods in recent years. An intuitive way is to segment the contours of the myocardium of the whole cardiac sequence first and then derive the motion status based on these contours. However, obtaining the myocardium contours itself is another hot topic in cardiac image analysis and requires a large amount of manually contoured cardiac sequence for model learning. Therefore, to alleviate the challenges of the four-way motion scoring, most of the existing methods instead concentrate on abnormality detection.

In summary, these methods follow a pipeline of: 1) myocardium segments localization, 2) motion information extraction, and 3) motion classification. They can be grouped into two categories, i.e, statistics representation-based methods, and shape model-based ones. The first category captures the complex dynamic information by extracting the statistical representation. In early work of [13], the local myocardial contraction pattern was extracted by independent component analysis. Latter in [15], [16], the myocardial points were estimated by an unscented Kalman smoother and were used to measure the information of left ventricle (LV), including radial distance, segment area, arc length, wall thickness, and radial velocity. The information was further explored to detect the abnormality by Shannon's differential entropy [7], [14]. Then, more statistical features have been explored. The proportion of blood within each segment was used to characterize the segmental contraction of the myocardium [4], [11], thus avoiding LV boundary delineating in all the frames. Parameters extracted from the time-signal intensity curves in radial spatial-temporal image profiles were used to assess wall motion in LV function by using kernel dictionary learning [6].

The second category tends to capture the abnormal region information by shape analysis. The principal component analysis and orthomax rotations were used to build a sparse localized shape model of LV [5]. A novel tensor-based classification framework [18] was used to identify and localize regional abnormal cardiac function by the myocardial strain pattern from tagged MRI, which was extracted by a non-tracking-based strain estimation method [19]. A wall segment model was developed for normal and abnormal hearts, and a Hidden Markov Model (HMM) was employed for classification [12]. In the work of [10], the relationships of spatiotemporal inter-landmark were used for shape extraction and interpretation. In the work of [8], a differentiable manifold was used to explore the most correlated regions.

While these methods achieved promising accuracy of abnormality detection, they are still incapable for the task of motion scoring because: 1) only heuristic handcrafted feature is not discriminative to capture the complex regional motion information; 2) delineation of myocardium border was required, either manually or semi-automatically, introducing inconvenience in clinical practice; 3) they only focused on abnormal motion detection, not the four-way motion scoring task. The proposed method learns discriminative motion features by the novel MMSA mechanism in multi-scale space and avoids the requirements of myocardium contours.

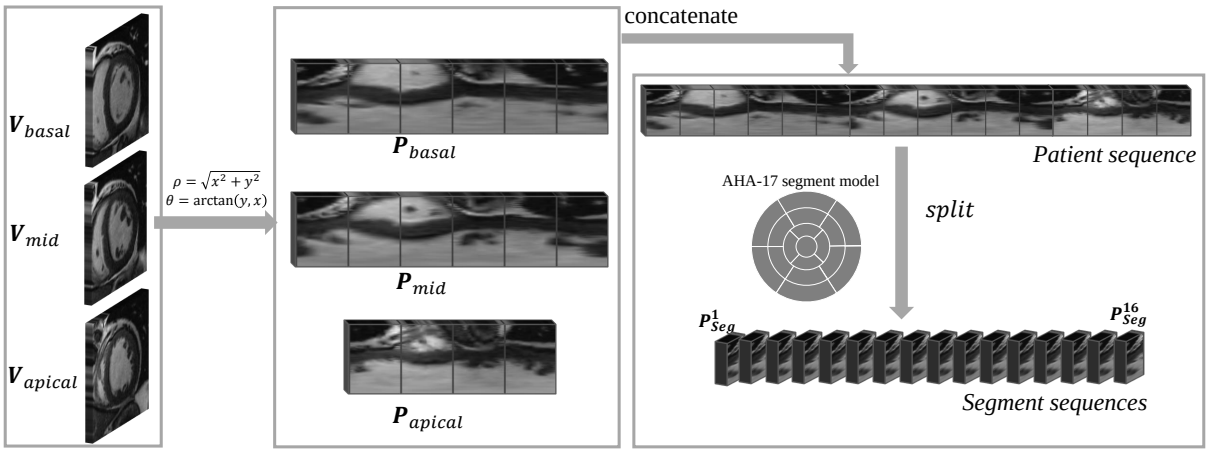


Fig. 1: Preprocessing of cardiac sequences, which includes conversion of the sequences into the polar space that makes the myocardium wall motion easier to detect, and rearrangement into patient sequences for motion extraction and into segment sequences for motion scoring according to the AHA 17-segment heart model [1].

B. Motion information extraction

Effective extraction of motion information from image sequences plays an important role for a wide range of applications on video parsing, as well as the task of cardiac motion scoring. Generally, motion information can be utilized 1) either as a complementary to the appearance model to maintain the temporal consistency [20]–[22], or 2) as an indicator to help identify the moving objects [23]. In the work of [20], motion and appearance features were combined within a two-stream CNN to help achieve accurate segmentation of generic objects in videos. Motion information was explicitly learned by optical flow estimation and propagated to the segmentation branch to help maintain the motion consistency for video object segmentation in [21], and was implicitly extracted by the task of future frame prediction to guide the segmentation module focusing on object regions for accurate results in [23].

Motion information was also explored in cardiac sequences analysis [22], [24]–[27]. Feature embeddings of optical flow between successive frames were concatenated with those of echocardiogram for segmentation of echocardiography [24], where the optical flow was computed in prior by the Horn-Schunck algorithm. Co-learning of segmentation and optical flow was used in [27] for echocardiography segmentation. A localized anatomical constrained optical flow estimation method was proposed in [25] for left ventricle tracking of 3D echocardiography. An optical flow network was embedded in the encoder of UNet in addition to the appearance of left ventricle segmentation from cardiac cine MRI [22]. Optical flow estimation was jointly optimized with cardiac segmentation using two parallel branches to improve each other.

In these methods, the optical flow was either used to make the extracted features more related to the movement of the myocardium or as complementary information for the main task. Instead, we propose to utilize the motion information 1) as a top-down mechanism to regularize the convolution feature learning, and 2) as an attention module to guide the identification of important regions in the bottom-up feature extraction procedure in multi-scale space.

III. Methodology

TABLE I: The legend of abbreviations and acronyms.

MMSA	**M**ulti-sacle **M**otion-based **S**patial **A**ttention
CNN	**C**onvolutional **N**erual **N**etworks
FC	**F**ully-**C**onnected layer
TSE	**T**emporal **S**queeze and **E**xcitation module
NL	**N**on**L**ocal block
AD	**A**bnormal **D**etection
MSI	**M**otion **S**core **I**ndex

Given the existing limitations in cardiac motion analysis methods, we propose a deep neural network that can 1) estimate the multi-scale motion information with an unsupervised top-down optical flow branch, and 2) predict the status of myocardium segment motion with a supervised bottom-up feature extraction branch, where a multi-scale motion-based spatial attention (MMSA) module is designed to help extract motion-aware features from the low-level convolution features. The two branches interact on multiple scales, rather than flow in parallel in existing methods. In this section, we first describe the preprocessing procedure of the cardiac MR sequences. Then the architecture of the MMSA-based motion scoring is introduced, and the MSA module is detailed. Further enhancements of the obtained features are introduced next to explore the dependencies between spatial-distant regions and temporal-distant frames.

A. Pre-processing of cardiac sequences

To alleviate the difficulty of myocardium motion extraction from cardiac sequences, all short-axis cardiac cine sequences are first converted to the polar space, and then rearranged into patient sequences and segment sequences according to the AHA-17 segment model of the left ventricle [1]. Fig. 1 shows the pre-processing procedure.

Firstly, all cardiac sequences, including the basal, middle, and apical slices, of each patient are converted to the polar space to make the motion extraction more efficient. The myocardium of the left ventricle forms an approximately

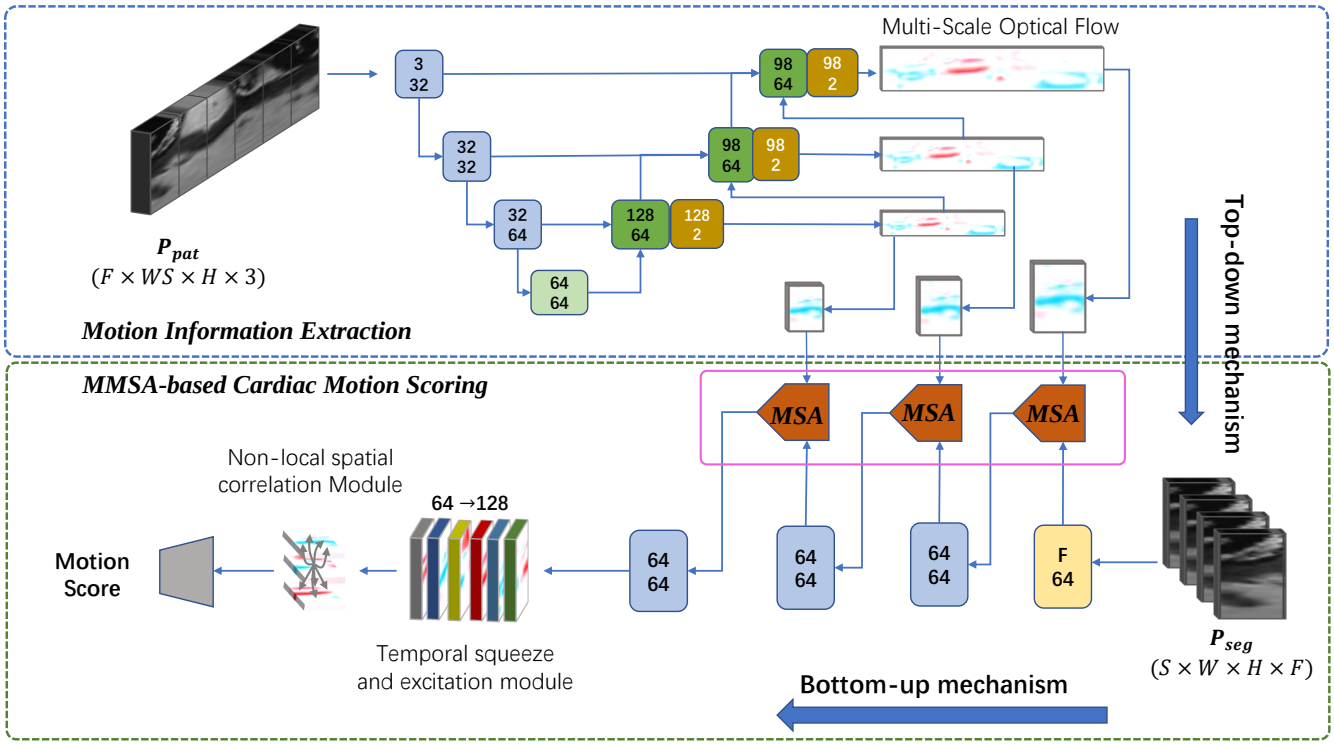


Fig. 2: The architecture of our MMSA-based cardiac motion scoring, which is constituted of a motion information extraction part (as shown in the blue rectangle) and a motion scoring part (as shown in the green rectangle). The bottom-up feature extraction in the motion scoring part is guided by the motion information through a newly proposed multi-scale motion-based spatial attention (MMSA) module in a top-down mechanism.

circular shape, and the motion of the myocardium can be mainly described by 1) wall thickening, 2) myocardium diastole/systole, and 3) circumferential movement. It is more efficient to capture the first two types of movement along the vertical axis and the last one along the horizontal axis in the polar space, as shown in Fig. 1. Besides, the motion of other moving objects, such as papillary muscles and trabeculations, can also be identified from sequences of polar space [9].

The conversion can be described by Eq. 1:

$$[\mathbf{P}_{basal}, \mathbf{P}_{mid}, \mathbf{P}_{apical}] \xleftarrow[\theta=arctan(y,x)]{\rho=\sqrt{x^2+y^2}} [\mathbf{V}_{basal}, \mathbf{V}_{mid}, \mathbf{V}_{apical}] \tag{1}$$

where $\mathbf{V}_{basal}, \mathbf{V}_{mid}, \mathbf{V}_{apical}$ denote the original cardiac sequences of three difference slices, corresponding to the basal, middle, and apical of LV. $\mathbf{P}_{basal}, \mathbf{P}_{mid}, \mathbf{P}_{apical}$ denote their counterparts in the polar space, and each $\mathbf{P}_* \in R^{r \times a \times t}$. r, a are the sampling numbers along the radial and angular dimension, and t is the frame number in the cardiac sequence.

Secondly, all sequences of each patient in the polar space are accommodated to patient-level sequence $\mathbf{P}_{pat}$ and segment-level sequences $\mathbf{P}^k_{seg}, k = 1, ..., 16$, for the tasks of global motion evaluation and regional segment-level motion scoring, respectively. For global evaluation, to accurately capture both the vertical movements within each segment and the horizontal movements that may cross neighboring segments, the patient-level sequence is fed into the optical-flow branch. For motion scoring, to obtain detailed segment-level motion scores, the segment-level sequences are fed into the feature extraction branch. The accommodation procedure is conducted according to the AHA-17 segment model and is shown in Fig. 1. Since there are only four segments in the apical slices, the apical sequence is first resampled along the angular dimension so that all the 16 segments have the same dimension. $\mathbf{P}_{pat}$ and $\mathbf{P}^k_{seg}$ can be obtained by the following equations:

$$\begin{aligned} [\mathbf{P}^1_{seg}, ..., \mathbf{P}^6_{seg}] &= split(\mathbf{P}_{basal}, 6) \\ [\mathbf{P}^7_{seg}, ..., \mathbf{P}^{12}_{seg}] &= split(\mathbf{P}_{mid}, 6) \\ [\mathbf{P}^{13}_{seg}, ..., \mathbf{P}^{16}_{seg}] &= split(\mathbf{P}_{apical}, 4) \end{aligned} \tag{2}$$

$$\mathbf{P}_{pat} = [\mathbf{P}^1_{seg}, ..., \mathbf{P}^{16}_{seg}] \tag{3}$$

where $split(\mathbf{P}, n)$ means splitting $\mathbf{P}$ into n parts of equal size along the first dimension.

B. MMSA-based cardiac motion scoring

The architecture of the MMSA-based cardiac motion scoring is demonstrated in Fig. 2. A bottom-up convolution neural network (CNN) is used to learn discriminative features for each myocardium segment, while a top-down optical flow estimation network is employed to complement the spatial CNN features with motion information and help identify important regions in the CNN feature maps by a newly designed multi-scale motion-based spatial attention (MMSA) module. The optical flow branch works in an unsupervised manner and in multi-scale. The extracted motion information then improve

the CNN branch from two aspects: 1) acts as complementary information to the CNN feature by directly concatenating motion and spatial features of the same scale; and 2) acts as a top-down attention mechanism to identify important regions of the CNN features, therefore guide the learning procedure of the CNN features. The proposed MMSA module plays a critical role in connecting the motion information and the CNN feature, and will be detailed in III-C.

The bottom-up CNN branch takes the segment sequences $\mathbf{P}_{seg}^{k}$ as input and learns multi-scale hierarchical features for the task of motion scoring, and the ground truth motion score of each segment is used during learning. In this work, we used the similar network architecture of [17] (see the green rectangle of Fig. 2), which contains four successive convolution blocks and two fully-connected (FC) layers for final score prediction. Each block contains two convolution layers and one down-sampling layer. The leaky-Relu layer is used as the nonlinear activation layer after each convolution. Specifically, the kernel interpolation convolution layer proposed in [17] is used in the first convolution layer to adapt cardiac sequences with different lengths. The output convolution features of each block are denoted as $\mathbf{x}^i, i = 1, 2, 3$. (For simplicity, the index of the segment is ignored here.) Besides, the extracted features are further enhanced by a spatial non-local correlation module and by a temporal squeeze-excitation module in the temporal domain. This enhancement will be detailed in III-D.

The top-down optical flow branch takes the patient sequence $\mathbf{P}_{pat}$ as input and learns the motion information between adjacent frames in an unsupervised and multi-scale manner. Inspired by the work of end-to-end optical flow estimation [28], we deploy an encoder-decoder architecture for cardiac motion estimation. The convolution blocks in the encoder and decoder are similar to those in the CNN branch, except that upsampling rather than downsampling is used in the decoder. Skip connection is utilized between the encoder and the decoder to keep more spatial information of the cardiac structure during the optical flow estimation in the decoder part. Different from the work in [28], each block in the decoder takes as input three types of signals: feature maps from the previous block, feature maps from skip connection, and the predicted optical flow of the previous block, and give two outputs: the upsampled feature maps and the optical flow prediction in the current scale. Details can be seen in the blue rectangle of Fig. 2. The higher block thus learns high-resolution results from the prediction of a lower block. In such a way, multi-scale motion information of the myocardium can be obtained. This also enables us to build multi-scale connections (see MMSA in III-C) between feature maps of the CNN branch and motion information of the optical flow branch. Each block of the newly designed decoder can be formulated as:

$$\mathbf{e}_{de}^{i} = convblock([\mathbf{e}_{en}^{i-1}, \mathbf{e}_{de}^{i-1}, \mathbf{u}^{i-1}]) \tag{4}$$

$$\mathbf{u}^{i} = conv(\mathbf{e}_{de}^{i}) \tag{5}$$

where $\mathbf{u} = [u_v, u_h]$ is the estimation of optical flow along two directions, and $\mathbf{e}_{en}^{i}, \mathbf{e}_{de}^{i}$ are the feature maps from the ith block in the encoder and the decoder, respectively.

Given the assumption of brightness consistency and spatial smoothness, we construct the following photometric objective function to extract motion information:

$$L_{con,k}^{i} = \|P_{pat}^{i} - warp(\mathrm{TS}(P_{pat}^{i}, d), \mathbf{u}^{i})\|^2 \tag{6}$$

$$L_{smo}^{i} = \| \bigtriangledown u_v\|^2 + \| \bigtriangledown u_h\|^2 \tag{7}$$

$$L_{optical}^{i} = \frac{\lambda_1}{2}(L_{con,d}^{i} + L_{con,-d}^{i}) + \lambda_2 L_{smo}^{i} \tag{8}$$

$i = 1, 2, 3$ denotes the different scale of the optical filed, as shown in Fig. 2. $\mathrm{TS}(P, k)$ means circular shift of P by k frames along the temporal dimension. $warp(P, \mathbf{u})$ means deforming P with the vector field $\mathbf{u}$. Therefore, $L_{con,d}^{i}$ denotes the reconstruction error between the current sequence P and its d-time shift version. $\bigtriangledown$ is the Sobel gradient operator. λ_1, λ_2 are the weighting factors between the consistency term and the smoothness term.

The final multiscale objective function of the optical flow branch is then

$$L_{optical} = \sum_{i=1}^{3} L_{optical}^{i}. \tag{9}$$

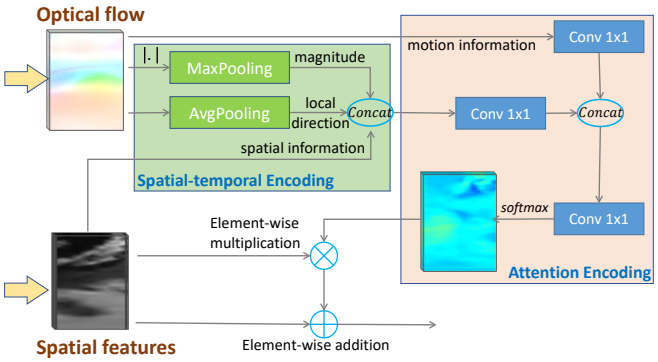


Fig. 3: The details of the Motion-based Spatial Attention Module (MSA).

C. Multi-scale Motion-based Spatial Attention

When the bottom-up spatial CNN features and the top-down temporal motion information of the myocardium are available, a crucial step for motion scoring is to effectively leverage the most relevant information from both of them. A frequently way used is concatenating them directly. Since the motion status of each myocardium segment is closely related to the segment itself, and the irrelevant background regions possess large areas, leading to low sensitivity to myocardium motion. Therefore, in this work, we design a motion-driven spatial attention module and deploy it between the CNN branch and the optical flow branch in multi-scales, as shown by the pink rectangles in Fig. 2. This Multi-scale Motion-based Spatial Attention mechanism is dubbed as MMSA. In such a way, important regions of the spatial feature can be identified, adaptively weighted and combined with the motion information.

With the CNN feature $\mathbf{x}^i$ and the motion information of the corresponding scale $\mathbf{u}^i$, we design the motion-based spatial attention module with two encodings, i.e., *spatial-temporal encoding* and *attention encoding*, as illustrated in Fig. 3. Since

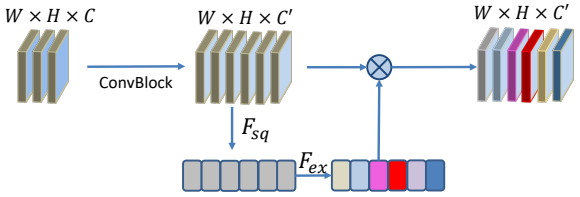


Fig. 4: The temporal squeeze and excitation (TSE) module adaptively weight the features along the temporal dimension.

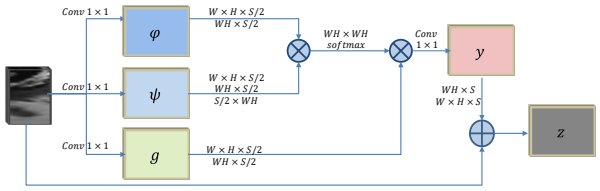


Fig. 5: The spatial non-local module explore long-range dependency of local responses across all locations in the feature map, thus combines local and distant relevant responses for the final motion scoring.

the motion scoring is operated at segment-level. The patient-level motion information is first split into segments, i.e, the inverse operation of Eq. 3.

$$[\mathbf{m}_1^i, ..., \mathbf{m}_{16}^i] = split(\mathbf{u}^i), i = 1, ..., 3 \tag{10}$$

We ignore the index for the segment and scale hereafter for simplicity.

In *spatial-temporal encoding*, the motion features and the spatial features are first combined by concatenating the motion magnitude, local direction, and the spatial features as follows:

$$\mathbf{c}^i = concat[maxpool(|\mathbf{m}^i|), avgpool(\mathbf{m}^i), \mathbf{x}^i]. \tag{11}$$

In *attention encoding*, the obtained spatial-temporal features $\mathbf{c}^i$ and the original optical flow of all frames are first aggregated with a 1×1 convolution kernel along the temporal dimension. The results are concatenated together and applied with another 1×1 convolution, leading to a feature map $\mathbf{M}$ with one channel. High values in $\mathbf{M}$ can be utilized to localize important motion regions of the myocardium segment. We normalize it with a softmax operation:

$$\mathbf{A}(p, q) = \frac{\exp\left(\mathbf{M}(p, q)\right)}{\sum_p \sum_q \exp\left(\mathbf{M}(p, q)\right)} \tag{12}$$

where p, q denote the coordinates in spatial domain.

Then, the attention map $\mathbf{A}$ can be used to allocate different weights to the spatial CNN features $\mathbf{x}^i$, thus highlighting the motion-relevant regions and diminishing the irrelevant background regions. To ease the learning of the attention module with a warm start, we add the CNN feature map $\mathbf{x}^i$ to the attention-weighted results, similar to the skip connection in ResNet [29]. The output of the MSA module can be formulated as:

$$\mathbf{z}^i = \mathbf{A} \otimes \mathbf{x}^i + \mathbf{x}^i \tag{13}$$

where $\otimes$ denotes element-wise multiplication with $\mathbf{A}$ broadcasting along the channel dimension.

D. Spatial and temporal feature enhancement

To further improve the temporal and spatial discrimination of the features before the final segment-sequence scoring, two feature enhancement modules are introduced.

Firstly, a temporal squeeze and excitation (TSE) module is implemented to emphasize the role of key frames and weaken the effect of redundant frames. We implement this by the SE-module [30], as shown in Fig. 4. The input feature maps were first squeezed to $\mathbf{s}_1$ by a global average pooling F_{sq}, and then excited by two fully connected layers F_{ex}, one for dimension reduction and the other increment. The resulted vector $\mathbf{s}_2$ acts as a weighting factor for each channel, i.e., each frame in the sequence, resulting in the output x_{tse}. The procedure of TSE can be formulated as:

$$\mathbf{s_1} = F_{sq}(\mathbf{x}) = \frac{1}{H \times W} \sum_i \sum_j \mathbf{x}\left(i, j\right), \tag{14}$$

$$\mathbf{s_2} = F_{ex}\left(\mathbf{s}_1, \mathbf{W}_1, \mathbf{W}_2\right) = \sigma\left(\mathbf{W}_2 \delta\left(\mathbf{W}_1 \mathbf{s}_1\right)\right), \tag{15}$$

$$\mathbf{x_{tse}} = \mathbf{s}_2 \otimes \mathbf{x}, \tag{16}$$

where $\mathbf{W_{1/2}}$ denotes the weights of the two fully connected layers, δ is the leaky ReLU function with the leaky value of 0.2 in our case, and σ is the softmax function. As a result, the temporal responses can be adaptively re-calibrated and become more informative.

Secondly, the non-local (NL) spatial correlation module utilized in the previous work [17] is also used to explore latent dependencies and capture subtle differences among motion patterns of different regions. It computes the response at a position as a weighted sum of the features at distant related positions:

$$\mathbf{r}(i) = \frac{1}{\mathcal{C}(\mathbf{x})} \sum_{\forall j} h(\mathbf{x}(i), \mathbf{x}(j)) g(\mathbf{x}(j)), \tag{17}$$

where the pairwise function h computes the similarity between two positions i and j, g computes a representation of the input feature, and $\mathcal{C}(\mathbf{x}) = \sum_{\forall j} h(\mathbf{x}(i), \mathbf{x}(j))$ is a normalization factor. In this work, we use the embedded Gaussian [31] for h:

$$h(\mathbf{x}(i), \mathbf{x}(j)) = \exp^{\phi(\mathbf{x}(i))^T \psi(\mathbf{x}(j))}, \tag{18}$$

which makes $\frac{1}{\mathcal{C}(\mathbf{x})} h(\mathbf{x}(i), \mathbf{x}(j))$ as a softmax function. The computation of an NL block in neural network is shown in Fig. 5:

$$\mathbf{x}_{nl}(i) = \theta(\mathbf{r}(i)) + \mathbf{x}(i), \tag{19}$$

where the residual connection $+\mathbf{x}(i)$ allows a non-local block to be inserted into any pre-trained model, while keeping its initial behavior. ϕ, ψ and θ are implemented as convolution with 1×1 kernel.

Finally, a fully connection block with the softmax layer to predict the final motion score. Since the goal of the cardiac motion scoring task is a four-way classification task in essence, following the scoring system of [1], thus the objective function

to train the network is the Categorical Cross-Entropy loss:

$$Loss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} y_{i,j}\log\hat{y}_{i,j}, \tag{20}$$

where $\hat{y}_{i,j}$ and $y_{i,j}$ denote the predicted probability and the true probability of jth (i = 1,...16) segment sequence for patient i, respectively. N and K are the numbers of patients and segments, respectively.

IV. Dataset and Configuration

In this section, details of the cardiac MR dataset we used for validation of the proposed method are first introduced. Then the implementation details of the method are described. The evaluation protocol is also given in this section.

A. Dataset

To verify the effectiveness of our proposed work, cardiac MR sequences of 90 subjects (5775 images) are collected, which were first used in our preliminary work [17]. The dataset contains short-axis cine sequences from scanners of multiple vendors, with spatial resolutions of 0.6445∼1.9792 mm/pixel, the frame number of 20 (65 subjects) or 25 (25 subjects) in one cardiac cycle. Various pathologies are in the presence, including infraction, dilated cardiomyopathy, ventricular hypertrophy, ischemia, scar formation, *etc*. For each subject, three representative short-axis cine sequences are chosen from the basal, mid, and apical left ventricle. Segment-wise ground truth of motion score is obtained from the radiologists' report and experts' visual inspection, resulting in a total of 1440 segments in the dataset. The distribution of motion score is {794, 348, 207, 91} for the four motion categories, showing its challenging unbalance character.

For each sequence, three landmarks, i.e, the junctions of left and right ventricles, and the center point of the cavity, are manually pointed to crop the left ventricle and to align its orientation. Due to the different spatial resolution of MR images, the size of the cropped image ranges approximately from 60 to 160. For CMR images with a low spatial resolution (e.g., 1.9792mm/pixel), the ROI size is about 60. For those with high resolution (e.g., 0.6445mm/pixel), the ROI size is about 160. To avoid loss of image details, we resized all the cropped images to 160×160. Normalization of the intensity is conducted by contrast limited adaptive histogram equalization in Matlab with the following parameters: numtiles of [8, 8], cliplimit of 0.001 and Rayleigh distribution.

B. Experimental Configuration

Three-fold cross-validation is utilized to evaluate the proposed method. Specifically, segment-sequences from 60 subjects are randomly selected as the training dataset, and the rest are used as the test dataset. We repeat this procedure three times and report the results from all subjects. For the splitting, the prevalence of each motion type is guaranteed to be close for the training set and the test set, as Fig. 6 shows.

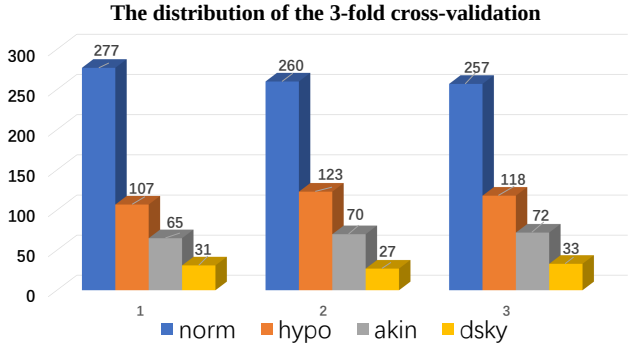


Fig. 6: The distribution of 3-fold cross-validation sets in our setting.

The network is trained using stochastic gradient descent with a batch size of 16, corresponding to the 16 segment-level sequences from one patient. To determine the hyper-parameters (such as the initial learning rate, weight decay, and momentum), we randomly chose 5 subjects as validation set from the training set of the first-round experiment. When these hyperparameters were determined, subjects of the whole training set were used for training. These hyperparameters were applied directly for experiments of the second and third rounds without further tuning. Finally, we use a weight decay of 0.1 with a momentum of 0.9995 and set the initial learning rate to 0.001. The learning rate is divided by 4 when the loss does not decrease. This setting is used to train a baseline model where no MMSA or nonlocal module is employed. During the finetuning phase, the initial learning rate is set to 0.0001 and is divided by 2 for every 1000 iterations with a total of 6000 iterations.

C. Evaluation Metrics

The performance of the proposed method is evaluated from three aspects. First, classification accuracy is used to evaluate the performance of the four-way classification motion scoring task.

$$acc_{ms} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{K}\mathbf{1}(y_{i,j} == \hat{y}_{i,j})}{N\times K}\times 100\% \tag{21}$$

Recall and precision are also computed for each of the motion type to demonstrate the specific performance for each type.

Second, the performance of the proposed method is evaluated for the task of cardiac motion score index (MSI) estimation. MSI is a continuous variable and is an important clinical factor for evaluation of the global cardiac dynamical function. It can be calculated as the average of the scores of all segments for each subject [1]. To evaluate the prediction consistency between the estimated results and the ground truth value, the Pearson correlation coefficient is used:

$$\rho_{msi} = corr(\frac{1}{K}\sum_{j=1}^{K} y_{i,j}, \frac{1}{K}\sum_{j=1}^{K}\hat{y}_{i,j}) \tag{22}$$

Third, to compare with the existing motion abnormality detection (AD) methods [4], [5], [7], classification accuracy

(acc_{ad}), Kappa value (κ_{ad}, which is a more convincing index considering the prevalence of positive and negative samples [32]), are calculated. For the proposed method, we binarize the ground truth and the prediction of motion score into normal and abnormal, and then compute acc_{ad} and κ_{ad} as follows:

$$acc_{ad} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{K}\mathbf{1}(y_{i,j}^{b} == \hat{y}_{i,j}^{b})}{N \times K} \times 100\% \quad (23)$$

$$\kappa_{ad} = \frac{acc_{ad} - r_0}{1 - r_0} \quad (24)$$

where y^b and $\hat{y}^b$ are the binarized ground truth and predictions for motion abnormality, and r_0 is the expected accuracy by random guess [32]. Besides, recall and precision are also demonstrated.

V. Results Analysis

To validate the performance of MMSA and demonstrate its advantages over existing works on cardiac motion analysis, we first examine its prediction performance on the dataset with 90 subjects, explore the role of each module, and then make comparisons with reported results of existing work.

A. Performance analysis of MMSA

Experimental results on the database show that our MMSA is capable of providing accurate analysis of global and regional myocardium motion function. Table II, shows the performance of MMSA for motion scoring, motion index estimation, and abnormality detection. Overall, MMSA delivers good results for cardiac motion analysis. It achieves an accuracy of 79.3% for the four-way classification motion scoring, 89.0% for the binary abnormality detection, and a correlation of 0.943 for the global motion score index estimation. Moreover, it achieves a recall of 79.9% and precision of 94.7% for abnormality detection, showing the ability of MMSA to identify abnormal wall motion. For each abnormal motion types, MMSA achieves low sensitivity (59.3%∼72.9%). On one side, this may be attributed to the low prevalence of these patterns in patients of the dataset. On the other side, this indicates that the identifying the differences between these patterns are of great challenge.

Fig. 7 demonstrates in bulleye view the regional motions scores of four subjects from the ground truth labels and our predictions. With the cardiac SAX cine sequence, our method is capable of accurately predicting the regional motion status of the myocardium segment. For subject 1, two segments, i.e, apical anterior and mid-ventricle infer-lateral regions, are misclassified from *hypokinetic* to *akinetic*. For subject 2, the normal apical septal region is falsely predicted as *dyskinetic*. For Subject 3, anterior and lateral regions of the mid-ventricle are falsely predicted. The last column shows the worst case, for which 7 out of 16 segments are falsely predicted. The differentiation between normal and dyskinetic motion patterns, in this case, is very difficult.

TABLE II: Performance of our proposed MMSA for motion scoring, motion index estimation and abnormality detection.

Motion scoring (4-way classification)				
	acc_{ms}	recall	precision	
Norm	79.3	91.4	88.6	
Hypo		60.9	65.6	
Akin		72.9	70.9	
Dysk		59.3	63.5	
Abnormal detection (binary classification)				
	acc_{ad}	recall	precision	κ_{ad}
	89.0	79.9	94.7	0.776
Motion score index (ρ_{msi})				
0.943				

TABLE III: Ablation studies of our method in aspects of the frame distance in optical flow estimation, attention modules, and the single scale version of the MSA module. '1' means the original scale, while '1/2' and '1/4' mean the scales with half and 1/4 the size of the original scale. Performance improvement over our final model (the line in boldface with $d = 3$) is also displayed.

Configuration	acc_{ms}/Δ (%)	acc_{ad}/Δ (%)
distance		
d=1	79.2 / -0.1	88.9 /-0.1
d=3	**79.3 / 0.0**	**89.0 / 0.0**
d=5	79.2 / -0.1	89.0 / 0.0
d=7	78.9 / -0.3	88.7 / -0.3
attention modules with $d = 3$		
-MMSA	78.1 / -1.1	88.4 / -0.6
-TSE	78.0 / -1.3	88.3 / -0.7
-Nonlocal	78.3 / -1.0	88.7 / -0.3
single scale with $d = 3$		
1/4	78.7 / -0.6	88.8 / -0.2
1/2	78.1 / -1.2	88.5 / -0.5
1	78.7 / -0.6	88.9 / -0.1

Besides the bulleye view that provides a detailed illustration of myocardium motion, our method also predicts the global MSI with a high correlation coefficient of 0.943. Fig. 8 shows the scatter plot and the Bland-Altman plot of our method for estimation of MSI. It can be drawn that our MMSA gives consistent good predictions of global cardiac motion function for subjects with either normal myocardium motion or mild to severe motion abnormality. For reference, the reported inter-observer correlation for MSI is 0.85 [33] for echocardiography. Therefore, the proposed MMSA has great potential in the clinical practice of cardiac motion analysis. For the task of abnormal motion detection, our MMSA achieves high accuracy of 89.0%, and a Kappa value of 0.776 for all the 1440 myocardium segments.

B. Ablation study

We further conducted ablation studies to validate the effectiveness of our method, in aspects of the frame distance in optical flow estimation, attention modules, and the multi-scale mechanism. The results are demonstrated in Table III. From the first block, we can draw that in the optical flow estimation, $d = 3$ delivers the best performance, while smaller or large frame distance leads to slightly inferior performance. In the following experiments, we use $d = 3$ for our method.

From the second block, we can see the performance decreases when each of the attention modules is removed. For a fair comparison, when MMSA is removed, concatenation of

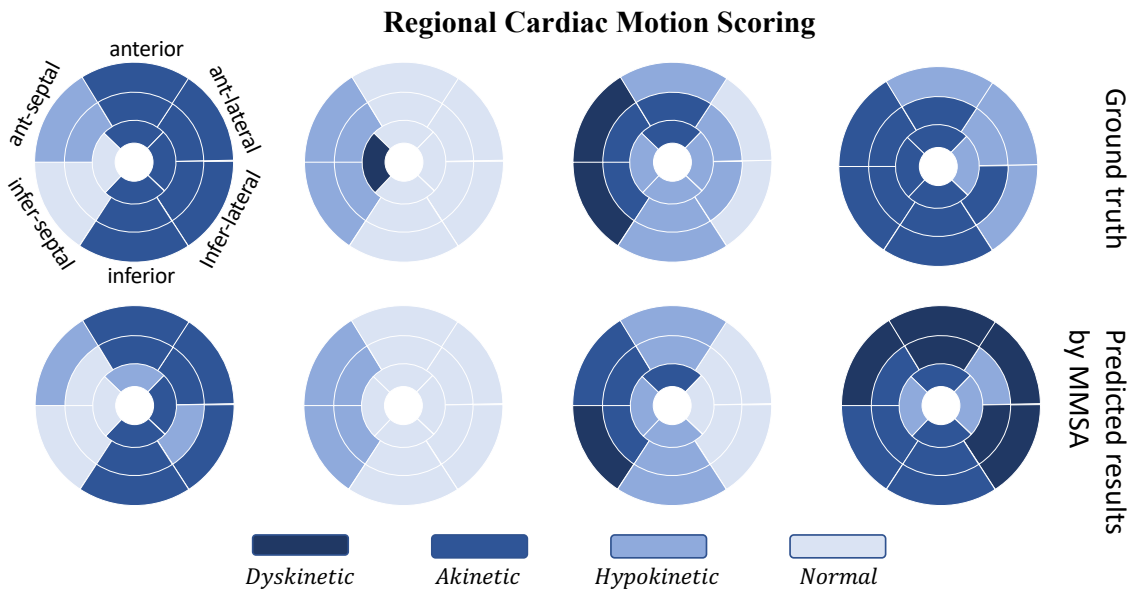


Fig. 7: Bull eye visualization of cardiac motion score for four subjects from the dataset. Our method is capable of giving accurate regional cardiac motion score from SAX cine sequence. The last column shows the worst case by our MMSA.

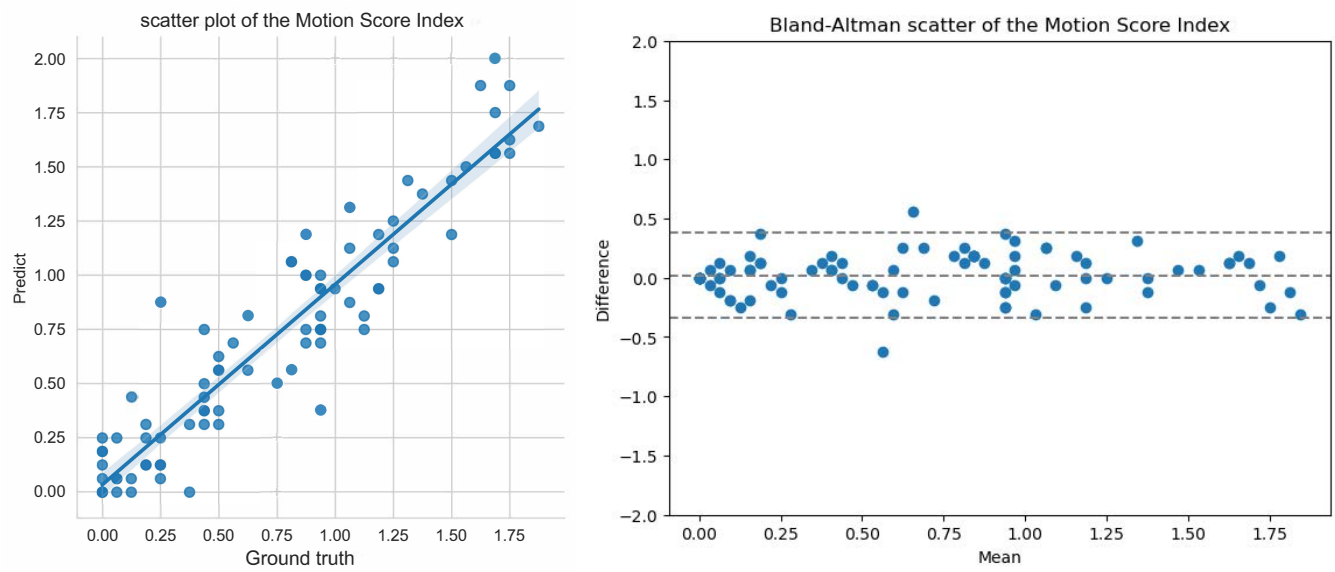


Fig. 8: The scatter plot (left) and the Bland-Altman plot (right) of the proposed method for MSI estimation. The horizontal dashed lines show the mean±1.96SD of the difference between estimations of MMSA and the ground truth.

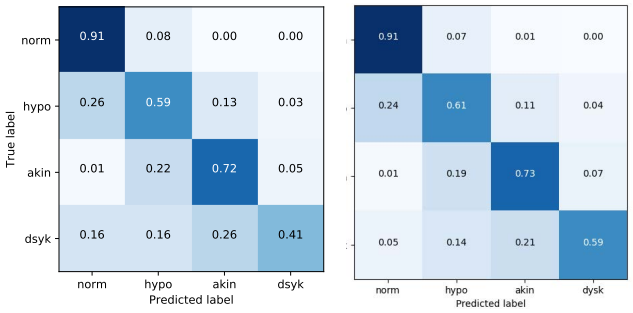
the CNN features and the optical flow is used as an alternative way of combining both. When TSE is removed, a convolution block is used instead to keep the network depth unchanged. It can be drawn that the absence of these modules results in inferior performance for all the tasks of motion scoring, and AD. When MMSA is removed, only concatenation cannot effectively exploit the motion information in the cardiac sequence. The multi-scale MSA modules can help to extract the most important features from the convolution feature maps. When TSE is removed, the accuracies of motion scoring decrease by 1.3% and 0.7% for AD. Adaptively weighting the extracted temporal features of the previous layer helps improve the final performance. When the NL module is removed, the accuracy of motion scoring decreases by 1.0% and 0.3% for AD. Long-range dependency captured by the NL module help differentiates motion patterns of subtle differences, as demonstrated in [17].

From the third block, we can draw that when deployed only on one scale, MSA cannot obtain the same accuracy for both tasks of motion scoring and AD. The multi-scale MSA can help extract the features that are most related to the myocardium motion in different scales, thus can be more sensitive to motions of various magnitude.

TABLE IV: Performance comparison

Method	# of Subject	Modality	Train/Test	acc_{ms} (%)	ρ_{ms}	acc_{ad} (%)	κ_{ad}
Information measure [15]	30	SAX MR	LOO	-	-	90.8	0.738
Registration and track [14]	30	SAX, 2C, 3C, 4C MR	LOO	-	-	91.9	0.738
Information approach [7]	58	SAX MR	LOO	-	-	87.1	0.73
Spatial-temporal tensor [19]	22	SAX Tagged MR	LOO	-	-	87.8	-
Correlation measure [9]	17	SAX Basal MR	No split	-	-	86.3	0.693
Statistical feature [4]	58	SAX MR	3-fold CV	-	-	86	0.73
Statistical shape model [13]	89	SAX MR	45/44	-	-	65.9	-
Shape Variations [5]	129	2C and 4C Echo	65/64	-	-	76.5	-
CNN	90	SAX MR	3-fold CV	74.7	0.913	85.8	0.711
Cardiac-MOS [17]	90	SAX MR	3-fold CV	77.4	0.926	87.8	0.749
MMSA	90	SAX MR	3-fold CV	**79.3**	**0.943**	**89.0**	**0.776**

SAX: short axis view; 2C/3C/4C: two/three/four chamber view; LOO: leave-one-subject-out; CV: cross validation



(a) Cardiac-MOS. (b) Our proposed MMSA.

Fig. 9: Confusion matrices of Cardiac-MOS and the proposed MMSA demonstrate that MMSA can reduce the misclassification of similar motion patterns (see the lower triangle of the matrices), therefor improves the correct prediction (see diagonal of the matrices).

C. Performance comparison with existing methods

To compare our MMSA with the existing methods of cardiac motion analysis, we summarized the reported results of these methods in Table IV. Due to the different experimental settings and datasets that were used in these methods, this table provides an intuitive comparison of these methods. We can draw from the table that: 1) the proposed MMSA achieves the best kappa value 0.776 for wall motion abnormality detection given the fact that our method is optimized with respect to the task of motion scoring instead of AD; 2) when compared with our preliminary work Cardiac-MOS, MMSA achieves improved performance for tasks of motion scoring, MSI estimation, and AD, consistently, which validates the effectiveness of the new MSA module for myocardium motion analysis.

Fig. 9 demonstrates the confusion matrices of Cardiac-MOS and MMSA for the task of motion scoring, respectively. From the low triangle of the two matrices, we can draw that MMSA can obviously reduce the misclassification of similar motion patterns. For example, the ratio of error reduces from 0.26 to 0.21 for misclassification of *dsyk* as *akin*, and from 0.16 to 0.05 for misclassification of *dysk* as *norm*. With the more effective MSA modules, slight differences between similar motion patterns of myocardium segments can be well captured, therefore leading to improved motion scoring performance.

VI. Discussion

Identifying the myocardial segments and extracting the motion features of the corresponding region are two indispensable factors for the task of motion scoring. The first factor usually can be achieved by delineation of the myocardium border either manually or semi-automatically, which results in additional workload and brings down the efficiency of clinical application. In our work, we alleviate this requirement by using the polar space to observe the myocardium movement and proposing the motion-based spatial attention module to highlight the myocardium regional. The second factor, i.e., the motion features, was usually calculated as hand-crafted features considering the temporal profiles of the cardiac sequence. We replace this with the optical flow estimation network, and optimize it together with the motion scoring task. Therefore, improved motion features that maximumly respond to the motion score can be obtained, and the four-way motion scoring task can be achieved.

Further efforts in cardiac motion analysis can be devoted to the following aspects. First, dense prediction of motion score could be investigated for every pixel-level location, so that more detailed identification of the motion status in the myocardium can be achieved, especially for the abnormal region. Secondly, motion features in the 3-D space could be explored to analyze the myocardium motion of any direction, including circumferential, longitudinal, and radial contraction/dilation. Our method focused only on movements of within-slice radial direction and ignored the tangential direction and out-of-plane movement. Thirdly, cardiac motion analysis can benefit a lot from the segmentation of epi- and endo- cardium border, which is a hot topic for cardiac image analysis and is usually used for ejection fraction estimation. The segmentation results not only help identify the myocardium region accurately in the images, but also contribute to strain analysis of the myocardium, which, together with motion scoring, are important for the diagnosis of coronary artery disease, heart failure, aortic stenosis, *etc*.

VII. Conclusion

In this paper, we propose an effective method MMSA for the task of myocardium motion scoring from short-axis cardiac MR sequences. In MMSA, a bottom-up low-level feature extraction mechanism and another top-down motion extraction

mechanism were connected through the multi-scale implementation of a novel motion-based spatial attention (MSA) module. The MSA module can effectively extract the most related low-level features according to the top-down guidance of motion extraction. Before the final prediction, the extracted features were further enhanced by a temporal squeeze-and-excitation module and a spatial non-local module. Experimental results validated the effectiveness of each module in MMSA, and revealed state-of-the-art performance for motion scoring, MSI estimation, and motion abnormality detection. Given the clinical significance of myocardium motion scoring, the proposed method has a great potential for practical assessment of cardiac motion function.

REFERENCES

[1] R. M. Lang, L. P. Badano, V. Mor-Avi, J. Afilalo, A. Armstrong, L. Ernande, F. A. Flachskampf, E. Foster, S. A. Goldstein, T. Kuznetsova *et al.*, "Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the american society of echocardiography and the european association of cardiovascular imaging," *European Heart Journal-Cardiovascular Imaging*, vol. 16, no. 3, pp. 233–271, 2015.

[2] E.-S. H. Ibrahim, "Myocardial tagging by cardiovascular magnetic resonance: evolution of techniques–pulse sequences, analysis algorithms, and applications," *Journal of Cardiovascular Magnetic Resonance*, vol. 13, no. 1, p. 36, 2011.

[3] I. Paetsch, C. Jahnke, V. A. Ferrari, F. E. Rademakers, P. A. Pellikka, W. G. Hundley, D. Poldermans, J. J. Bax, K. Wegscheider, E. Fleck *et al.*, "Determination of interobserver variability for identifying inducible left ventricular wall motion abnormalities during dobutamine stress magnetic resonance imaging," *European heart journal*, vol. 27, no. 12, pp. 1459–1464, 2006.

[4] M. Afshin, I. B. Ayed, K. Punithakumar, M. W. K. Law, A. Islam, A. Goela, T. M. Peters, and S. Li, "Regional assessment of cardiac left ventricular myocardial function via MRI statistical features," *IEEE TMI*, vol. 33, pp. 481–494, 2014.

[5] K. E. Leung and J. G. Bosch, "Localized shape variations for classifying wall motion in echocardiograms," in *MICCAI*. Springer, 2007, pp. 52–59.

[6] J. Mantilla, J. L. Paredes, J.-J. Bellanger, E. Donal, C. Leclercq, R. Medina, and M. Garreau, "Classification of LV wall motion in cardiac MRI using kernel dictionary learning with a parametric approach," *EMBC*, pp. 7292–7295, 2015.

[7] K. Punithakumar, I. B. Ayed, A. Islam, A. Goela, I. G. Ross, J. Chong, and S. Li, "Regional heart motion abnormality detection: An information theoretic approach," *Medical image analysis*, vol. 17, no. 3, pp. 311–324, 2013.

[8] J. Garcia-Barnes, D. Gil, L. Badiella, A. Hernández-Sabaté, F. Carreras, S. Pujades, and E. Martí, "A normalized framework for the design of feature spaces assessing the left ventricular function," *IEEE Transactions on Medical Imaging*, vol. 29, no. 3, pp. 733–745, 2010.

[9] Y. Lu, P. Radau, K. A. Connelly, A. Dick, and G. A. Wright, "Pattern recognition of abnormal left ventricle wall motion in cardiac MR," *MICCAI*, pp. 750–758, 2009.

[10] K. Lekadir, N. G. Keenan, D. J. Pennell, and G.-Z. Yang, "An inter-landmark approach to 4-d shape extraction and interpretation: application to myocardial motion assessment in mri," *IEEE Transactions on Medical Imaging*, vol. 30, no. 1, pp. 52–68, 2010.

[11] M. Afshin, I. B. Ayed, K. Punithakumar, M. W. Law, A. Islam, A. Goela, I. Ross, T. Peters, and S. Li, "Assessment of regional myocardial function via statistical features in MR images," in *MICCAI*. Springer, 2011, pp. 107–114.

[12] S. Mansor and J. A. Noble, "Local wall motion classification of stress echocardiography using a hidden markov model approach," in *ISBI*. IEEE, 2008, pp. 1295–1298.

[13] A. Suinesiaputra, A. F. Frangi, T. A. Kaandorp, H. J. Lamb, J. J. Bax, J. H. Reiber, and B. P. Lelieveldt, "Automated detection of regional wall motion abnormalities based on a statistical model applied to multislice short-axis cardiac MR images," *IEEE TMI*, vol. 28, no. 4, pp. 595–607, 2009.

[14] K. Punithakumar, I. B. Ayed, A. Islam, A. Goela, and S. Li, "Regional heart motion abnormality detection via multiview fusion," in *MICCAI*. Springer, 2012, pp. 527–534.

[15] K. Punithakumar, I. B. Ayed, A. Islam, I. G. Ross, and S. Li, "Regional heart motion abnormality detection via information measures and unscented kalman filtering," in *MICCAI*. Springer, 2010, pp. 409–417.

[16] K. Punithakumar, I. B. Ayed, I. G. Ross, A. Islam, J. Chong, and S. Li, "Detection of left ventricular motion abnormality via information measures and bayesian filtering," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 4, pp. 1106–1113, 2010.

[17] W. Xue, G. Brahm, S. Leung, O. Shmuilovich, and S. Li, "Cardiac motion scoring with segment-and subject-level non-local modeling," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 437–445.

[18] Z. Qian, Q. Liu, D. N. Metaxas, and L. Axel, "Identifying regional cardiac abnormalities from myocardial strains using spatio-temporal tensor analysis," in *MICCAI*. Springer, 2008, pp. 789–797.

[19] ——, "Identifying regional cardiac abnormalities from myocardial strains using nontracking-based strain estimation and spatio-temporal tensor analysis," *IEEE TMI*, vol. 30, no. 12, pp. 2017–2029, 2011.

[20] S. D. Jain, B. Xiong, and K. Grauman, "Fusionseg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos," in *2017 IEEE conference on computer vision and pattern recognition (CVPR)*. IEEE, 2017, pp. 2117–2126.

[21] J. Cheng, Y.-H. Tsai, S. Wang, and M.-H. Yang, "Segflow: Joint learning for video object segmentation and optical flow," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 686–695.

[22] W. Yan, Y. Wang, Z. Li, R. J. Van Der Geest, and Q. Tao, "Left ventricle segmentation via optical-flow-net from short-axis cine mri: preserving the temporal coherence of cardiac motion," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 613–621.

[23] K. Xu, L. Wen, G. Li, L. Bo, and Q. Huang, "Spatiotemporal cnn for video object segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1379–1388.

[24] M. H. Jafari, H. Girgis, Z. Liao, D. Behnami, A. Abdi, H. Vaseli, C. Luong, R. Rohling, K. Gin, T. Tsang *et al.*, "A unified framework integrating recurrent fully-convolutional networks and optical flow for segmentation of the left ventricle in echocardiography data," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2018, pp. 29–37.

[25] S. Queirós, J. L. Vilaça, P. Morais, J. C. Fonseca, J. D'hooge, and D. Barbosa, "Fast left ventricle tracking using localized anatomical affine optical flow," *International journal for numerical methods in biomedical engineering*, vol. 33, no. 11, p. e2871, 2017.

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

[27] H. Wei, H. Cao, Y. Cao, Y. Zhou, W. Xue, D. Ni, and S. Li, "Temporal-consistent segmentation of echocardiography with co-learning from appearance and shape," in *MICCAI*, 2020.

[28] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2462–2470.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[30] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[31] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7794–7803.

[32] A. J. Viera, J. M. Garrett *et al.*, "Understanding interobserver agreement: the kappa statistic," *Fam med*, vol. 37, no. 5, pp. 360–363, 2005.

[33] K. Bjørnstad, M. Al Amri, J. Lingamanaicker, I. Oqaili, and L. Hatle, "Interobserver and intraobserver variation for analysis of left ventricular wall motion at baseline and during low-and high-dose dobutamine stress echocardiography in patients with high prevalence of wall motion abnormalities at rest," *Journal of the American Society of Echocardiography*, vol. 9, no. 3, pp. 320–328, 1996.