2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI) | 978-1-6654-2923-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISBI52829.2022.9761481

# SEMANTIC-AWARE TEMPORAL CHANNEL-WISE ATTENTION FOR CARDIAC FUNCTION ASSESSMENT

*Guanqi Chen* *Guanbin Li**

School of Computer Science and Engineering, Sun Yat-Sen University, China

## ABSTRACT

Cardiac function assessment aims at predicting left ventricular ejection fraction (LVEF) given an echocardiogram video, which requests models to focus on the changes in the left ventricle during the cardiac cycle. How to assess cardiac function accurately and automatically from an echocardiogram video is a valuable topic in intelligent assisted healthcare. Existing video-based methods do not pay much attention to the left ventricular region, nor the left ventricular changes caused by motion. In this work, we propose a semi-supervised auxiliary learning paradigm with a left ventricular segmentation task, which contributes to the representation learning for the left ventricular region. To better model the importance of motion information, we introduce a temporal channel-wise attention (TCA) module to excite those channels used to describe motion. Furthermore, we reform the TCA module with semantic perception by taking the segmentation map of the left ventricle as input to focus on the motion patterns of the left ventricle. Finally, to reduce the difficulty of direct LVEF regression, we utilize an anchor-based classification and regression method to predict LVEF. Our approach achieves state-of-the-art performance on the Stanford dataset with an improvement of 0.22 MAE, 0.26 RMSE, and 1.9% $R^2$.

***Index Terms*—** Ultrasound video, attention mechanism, semi-supervised auxiliary learning.

## 1. INTRODUCTION

Cardiac function is the capability of the heart to meet the metabolic demands of the body, which has a significant impact on human health. If the cardiac function is slightly damaged, fatigue, palpitations, dyspnea, or angina pectoris may occur, and in severe cases, heart failure may occur. In recent years, there are more patients with cardiac dysfunction than ever before, which leads to heart dysfunction becoming a global health problem [1]. Therefore it is very important to timely detect and treat cardiac dysfunction, which relies on an accurate assessment of cardiac function.

Left ventricular ejection fraction (LVEF), the ratio of the stroke volume to the end-diastolic volume in the left ventricle, is a critical metric of cardiac function. It reveals the effectiveness of pumping into the systemic circulation. In recent years, it has aroused great interest to use deep learning techniques [2, 3] on echocardiography to estimate LVEF. [2] tried to use a 2D CNN to assess cardiac function based on the manually selected images at end-systole and end-diastole. However, this simple method had a substantial error compared to the human assessment of cardiac function. [3] attempted to predict LVEF based on the echocardiogram video. They employed a spatiotemporal convolutional network to directly regress LVEF. Then they utilized another convolutional neural network (CNN) to segment left ventricles, which was used to identify the cardiac cycles. And they aggregated the results of several cardiac cycles to obtain the final prediction. Compared with [2], the method proposed in [3] made a great progress, which might benefit from the motion information in the echocardiogram video. However, current video-based methods ignore the positive effect of the left ventricular segmentation task on the cardiac function assessment task. The cardiac function assessment task requires models to focus on the changes of the left ventricle during the cardiac cycle, while the segmentation task can make networks capture discriminative representation for the left ventricular region. Besides, the left ventricular masks generated by the segmentation network can serve as cues, which include the spatial information of the left ventricles, to benefit the cardiac function assessment task.

Motivated by the above observations, in this paper, we propose a semi-supervised auxiliary learning paradigm to jointly address cardiac function assessment and left ventricular segmentation tasks. In the paradigm, we utilize a shared encoder and two different branches to predict LVEF and segment left ventricles respectively, which can make the encoder capture a better representation of the left ventricle. To make full use of the motion information, we design a temporal channel-wise attention (TCA) module to excite those channels which are used to describe motion in the spatiotemporal features. Besides, with the help of left ventricular segmentation masks, we further reform the TCA module with semantic perception, which transforms the TCA module into semantic-aware temporal channel-wise attention (S-TCA) module. Through the S-TCA module, we can make our model focus on the motion patterns of the left ventricle. Moreover, inspired by anchor-based detection methods [4],

*Corresponding author.

Authorized licensed use limited to: The Chinese University of Hong Kong CUHK(Shenzhen). Downloaded on October 26,2025 at 11:46:38 UTC from IEEE Xplore. Restrictions apply.

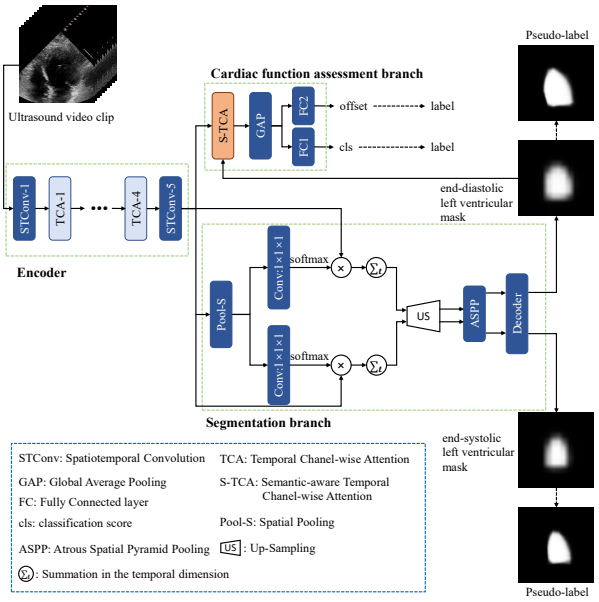


**Fig. 1**. Overview of our proposed method. Our model is composed of a shared encoder and two branches for left ventricular segmentation and cardiac function assessment respectively.

we introduce an anchor-based classification and regression method to predict LVEF instead of performing direct LVEF regression. In this way, we reduce the variance of the training samples, which is easier for the network to learn the potential expression. Finally, our approach achieves state-of-the-art performance on the Stanford dataset [3].

## 2. METHODOLOGY

The overall architecture of our proposed method is shown in Figure 1. A spatiotemporal convolution network with a set of temporal channel-wise attention (TCA) modules is employed to obtain the feature representation for the input video clip. And these TCA modules are designed to excite the motion patterns at a frame-level. After that, two branches are applied to address left ventricular segmentation and cardiac function assessment tasks, respectively. With the predicted left ventricular segmentation masks, we use a semantic-aware temporal channel-wise attention (S-TCA) module to bridge two branches and excite the motion patterns of the left ventricle. Finally, we utilize an anchor-based classification and regression method to predict LVEF. In this section, we will elaborate our proposed components.

### 2.1. Temporal Channel-wise Attention

Spatiotemporal features captured by spatiotemporal convolutions are commonly composed of static appearance features

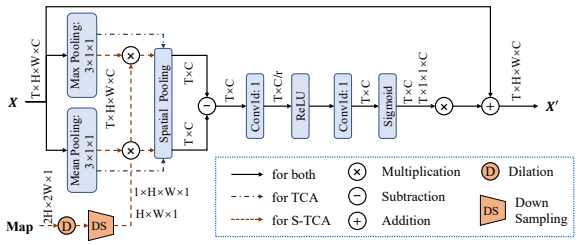


**Fig. 2**. Illustration for temporal channel-wise attention (TCA) module and semantic-aware temporal channel-wise attention (S-TCA) module.

and dynamic motion features. Due to the high frame rate and the fixed view of echocardiography, the appearance features of adjacent frames are similar. However, the motion patterns of the adjacent frames may be different, especially at end-systole and end-diastole. The direction of motions in these periods is the opposite of that in previous frames. Motivated by this observation, we argue that the appearance features after the max-pooling operation and the mean-pooling operation in the adjacent frames are the same, while the motion features may be different. Then we can leverage this difference to excite the motion features through a group of convolutions and activation functions.

The architecture of our proposed TCA module is illustrated in Figure 2. Given an input feature $X \in \mathbb{R}^{T\times H\times W\times C}$, local max-pooling and mean-pooling are applied to aggregate the information from adjacent frames and obtain features $X^{max}$, $X^{mean} \in \mathbb{R}^{T\times H\times W\times C}$. Then a global average pooling in the spatial dimension is adopted to $X^{max}$ and $X^{mean}$ to obtain global representations $\bar{X}^{max}$, $\bar{X}^{mean} \in \mathbb{R}^{T\times C}$. After that, based on the above analysis, we directly apply subtraction to global representations to stress motion features. Inspired by the SE block [5], we utilize two 1D convolutions with ReLU and sigmoid activation functions to obtain the temporal channel-wise attention weights $E \in \mathbb{R}^{T\times C}$.

$$E = sigmoid(W_2 * (ReLU(W_1 * (\bar{X}^{max} - \bar{X}^{mean})))), \quad (1)$$

where $*$ denotes the convolution operation. $W_1 \in \mathbb{R}^{C/r\times C\times 1}$ and $W_2 \in \mathbb{R}^{C\times C/r\times 1}$ are the parameters of 1D convolutions. $r = 16$ is the reduction ratio. Finally, we leverage $E$ to excite the motion patterns of $X$ in the following process:

$$X' = X \cdot E + X. \quad (2)$$

In this way, we can highlight the motion features without discarding the appearance features.

### 2.2. Semi-supervised Auxiliary Learning Paradigm

In order to make our model focus on the left ventricle, we introduce an auxiliary task to segment end-diastolic and end-systolic left ventricles. However, each video has only one

Authorized licensed use limited to: The Chinese University of Hong Kong CUHK(Shenzhen). Downloaded on October 26,2025 at 11:46:38 UTC from IEEE Xplore. Restrictions apply.

annotated cardiac cycle with two masks for the end-diastolic and end-systolic left ventricle, which can not ensure that there exist ground-truths for randomly sampled video clips. To address this issue, we utilize a DeeplabV3 [6] model trained on the sparse labels to generate the left ventricular masks for those unlabeled frames. Then we choose the predicted masks with the largest and smallest area among the input video clips as pseudo-labels to train the segmentation branch.

As shown in Figure 1, the segmentation branch contains two paths to segment the end-diastolic and end-systolic left ventricular masks. Next, we take how to obtain the end-diastolic left ventricular mask as an example to elaborate. Firstly, a spatial pooling is applied to the spatiotemporal feature $Z \in \mathbb{R}^{T\times H\times W\times C}$ to get the global representation $\bar{Z} \in \mathbb{R}^{T\times C}$. Then a 1D convolution and a softmax activation function are employed to $\bar{Z}$ to obtain the adaptive weight $R \in \mathbb{R}^{T\times 1}$ which represents the temporal relevance to the end-diastole.

$$R = softmax(W_3 * \bar{Z}), \tag{3}$$

where $W_3 \in \mathbb{R}^{1\times C\times 1}$ is the parameter of 1D convolution. After that, we obtain the end-diastolic representation $F \in \mathbb{R}^{H\times W\times C}$ through the following operation:

$$F = \sum_{t=1}^{T} Z_t \cdot R_t. \tag{4}$$

Then we upsample the representation $F$ for higher resolution, which is the same as the feature size in the DeeplabV3 model before the ASPP [6] module. Finally, we use an ASPP module and a decoder which consists of two convolutional layers to produce the end-diastolic left ventricular mask.

#### *2.2.1. Semantic-aware temporal channel-wise attention*

With the help of left ventricular masks, we can further reform the TCA module to focus on the motion patterns of the left ventricle. As shown in Figure 2, the semantic-aware temporal channel-wise attention (S-TCA) module utilizes the end-diastolic left ventricular mask to conceal the trivial region. Since sometimes the predicted mask is not accurate to cover the whole left ventricle, we expand the predicted area of left ventricle via dilation operation, which is achieved by the max-pooling operation.

### 2.3. Anchor-based Classification and Regression

Instead of directly regressing LVEF, we predict LVEF through two parts: classifying its interval and regressing the corresponding center offset. Firstly, we set $M = 20$ equally spaced anchor intervals, which cover the range of LVEF. Then we use two sibling fully connected layers following a global average pooling in spatiotemporal dimension to predict results. The first output is a vector of classification score for each anchor interval, and we can utilize a softmax function to turn it into a discrete probability distribution $p \in \mathbb{R}^M$. The second fully connected layer outputs a vector of offset $o \in \mathbb{R}^M$ for each anchor interval. And their corresponding labels, the ground-truth class $u \in \mathbb{R}^1$ and the ground-truth offset $v \in \mathbb{R}^M$, are obtained by following:

$$\begin{aligned} u &= \lfloor y/l \rfloor, \\ v_m &= \frac{y - c_m}{l}, \end{aligned} \tag{5}$$

where $y$ is the ground-truth of LVEF, and $l$ is the size of the anchor interval, which is equal to $100/M$. $m$ denotes the index of anchor intervals, and $c_m$ is the median value of the m-th anchor interval.

In this work, we utilize a multi-task loss function to train the whole model:

$$L = L_{ef} + \beta L_{aux}, \tag{6}$$

where $L_{ef}$ and $L_{aux}$ represent the loss of the cardiac function assessment task and the auxiliary task. $L_{ef}$ is consist of two parts, the cross-entropy loss of the anchor interval and the smooth $L_1$ loss [4] of the offset. And $L_{aux}$ is the cross-entropy loss between pseudo-labels and predicted left ventricular masks. $\beta$ is a hyperparameter for balancing two loss terms, which is set to 0.01 in our experiments since $L_{aux}$ is two orders of magnitude larger than $L_{ef}$.

## 3. EXPERIMENTS

In this section, we evaluate our proposed components on the Stanford dataset [3]. It is composed of 10030 labeled echocardiogram videos, including 7465 videos for training, 1288 videos for validation, and 1277 videos for testing. All videos are apical-4-chamber view, and they consist of a series of gray-scale images of $112 \times 112$ pixels.

### 3.1. Evaluation Metrics

In this paper, we only focus on the performance of the cardiac function assessment task, that is, the quality of the predicted LVEF. To quantitatively measure the model performance, we adopt three popular evaluation criteria: MAE, RMSE, and $R^2$.

### 3.2. Comparison with State-of-the-art Methods

In this section, we compare our proposed method with the existing state-of-the-art cardiac function assessment algorithms on the Stanford test set, which is shown in Table 1. EchoNet-Dynamic [3] uses a R(2+1)D model [7] to regress the LVEF, and employs the beat-to-beat evaluation strategy to assess the cardiac function. For fair comparisons, we utilize its open-source code to train several 3D CNNs which include MC3 [7], R3D [8], and R(2+1)D[7], in the same setting, then evaluate them with the same strategy of our model, which is to

Authorized licensed use limited to: The Chinese University of Hong Kong CUHK(Shenzhen). Downloaded on October 26,2025 at 11:46:38 UTC from IEEE Xplore. Restrictions apply.

**Table 1**. Comparison results with state-of-the-art algorithms on the Stanford test set. $^\star$ indicates the result is copied.

| Methods | MAE↓ | RMSE↓ | $R^2$ ↑ |
|---|---|---|---|
| MC3 [7] | 4.34 | 5.76 | 0.778 |
| R3D [8] | 4.16 | 5.55 | 0.794 |
| R(2+1)D [7] | 3.95 | 5.27 | 0.814 |
| EchoNet-Dynamic [3]$^\star$ | 4.05 | 5.32 | 0.81 |
| Ours | **3.83** | **5.06** | **0.829** |

**Table 2**. Ablation study on the Stanford test set. Sem represents semantic related strategies.

| Methods | ACR | TCA | Sem | MAE↓ | RMSE↓ | $R^2$ ↑ |
|---|---|---|---|---|---|---|
| $M_0$ | × | × | × | 3.95 | 5.27 | 0.814 |
| $M_1$ | ✓ | × | × | 3.92 | 5.21 | 0.819 |
| $M_2$ | ✓ | ✓ | × | 3.89 | 5.11 | 0.826 |
| $M_3$ | ✓ | ✓ | ✓ | **3.83** | **5.06** | **0.829** |

randomly sample 10 different clips from the video, and average the predictions of them to obtain the final prediction. As Table 1 displays, the proposed method achieves the lowest MAE and RMSE, the highest $R^2$ on the Stanford test set. Compared to the result reported in [3], our method considerably outperforms it by 0.22 MAE, 0.26 RMSE and 1.9% $R^2$.

### 3.3. Ablation Study

In Table 2, we verify the effectiveness of the components which constitute our proposed approach. $M_0$ represents the R(2+1)D model with a fully connected layer to regress LVEF directly. Based on $M_0$, $M_1$ employs the Anchor-based Classification and Regression method to predict LVEF rather than regress LVEF directly through a fully connected layer. Further, $M_2$ equips the R(2+1)D model with a set of TCA modules that follow the spatiotemporal convolution blocks. Finally, on the basis of $M_2$, $M_3$ is to make the model semantic-aware through adopting the semi-supervised auxiliary learning paradigm with the left ventricular segmentation task and replacing the last TCA module with the S-TCA module. As Table 2 indicates, each proposed component brings benefit to the complete method.

## 4. CONCLUSION

In this paper, we propose a new method for cardiac function assessment. A semi-supervised auxiliary learning paradigm is introduced to facilitate the representation learning for the left ventricular region. Besides, TCA and S-TCA modules are designed to excite those channels used to describe motion. Finally, in order to reduce the difficulty of direct LVEF regression, we utilize an anchor-based classification and regression method to predict LVEF. Our approach achieves state-of-the-art performance on the Stanford dataset.

## 5. COMPLIANCE WITH ETHICAL STANDARDS

This is a numerical simulation study for which no ethical approval was required.

## 6. ACKNOWLEDGMENTS

This work is supported by the Guangdong Basic and Applied Basic Research Foundation under Grant No.2020B1515020048, the National Natural Science Foundation of China (No.61976250), the Guangzhou Science and Technology Project (No.202102020633), and the Guangdong Provincial Key Laboratory of Big Data Computing, The Chinese University of Hong Kong, Shenzhen.

## 7. REFERENCES

[1] Boback Ziaeian and Gregg C Fonarow, "Epidemiology and aetiology of heart failure," *Nature Reviews Cardiology*, vol. 13, no. 6, pp. 368–378, 2016.

[2] Amirata Ghorbani, David Ouyang, Abubakar Abid, Bryan He, Jonathan H Chen, Robert A Harrington, David H Liang, Euan A Ashley, and James Y Zou, "Deep learning interpretation of echocardiograms," *NPJ digital medicine*, vol. 3, no. 1, pp. 1–10, 2020.

[3] David Ouyang, Bryan He, Amirata Ghorbani, Neal Yuan, Joseph Ebinger, Curtis P Langlotz, Paul A Heidenreich, Robert A Harrington, David H Liang, Euan A Ashley, et al., "Video-based ai for beat-to-beat assessment of cardiac function," *Nature*, vol. 580, no. 7802, pp. 252–256, 2020.

[4] Ross Girshick, "Fast r-cnn," in *ICCV*, 2015, pp. 1440–1448.

[5] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *CVPR*, 2018, pp. 7132–7141.

[6] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[7] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *CVPR*, 2018, pp. 6450–6459.

[8] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *ICCV*, 2015, pp. 4489–4497.

Authorized licensed use limited to: The Chinese University of Hong Kong CUHK(Shenzhen). Downloaded on October 26,2025 at 11:46:38 UTC from IEEE Xplore. Restrictions apply.