Cardiac MR Image Sequence Segmentation with Temporal Motion Encoding

Anonymous ECCV submission

Paper ID 10

Abstract. The segmentation of cardiac magnetic resonance (MR) images is a critical step for the accurate assessment of cardiac function and the diagnosis of cardiovascular diseases. In this work, we propose a novel segmentation method that is able to effectively leverage the temporal information in cardiac MR image sequences. Specifically, we construct a Temporal Aggregation Module (TAM) to incorporate the temporal image-based features into a backbone spatial segmentation network (such as a 2D U-Net) with negligible extra computation cost. In addition, we also introduce a novel Motion Encoding Module (MEM) to explicitly encode the motion features of the heart. Experimental results demonstrate that each of the two modules enables clear improvements upon the base spatial network, and their combination leads to further enhanced performance. The proposed method outperforms the previous methods significantly, demonstrating the effectiveness of our design.

Keywords: Cardiac MRI · Left ventricle segmentation · Temporal · Motion.

1 Introduction

Cardiac magnetic resonance imaging (MRI) is one of the major imaging modalities that can be used for the quantitative spatio-temporal analysis of cardiac function and disease diagnosis. Accurate assessment of cardiac function is essential for both diagnosis and treatment of cardiovascular diseases. Recent developments in machine learning methods promise to enable the design of automatic cardiac analysis tools, thereby significantly reducing the manual effort currently required by clinicians. In particular, the automatic segmentation of left ventricle (LV) contours is an important first step to enable the accurate quantification of regional cardiac function, including ejection fraction, temporal changes in ventricular volumes and strain analysis of the myocardium. However, accurate LV boundary segmentation is challenging due to LV shape variability, imaging artifacts, and poor LV boundary delineation. Such complexities make this task still an open problem despite the existence of important works for several decades.

Recent methods for cardiac LV segmentation are mainly based on deep neural networks, given their superior performance. One representative of the recent development is the 2D U-Net [8], which has proven one of the most effective methods in image-based segmentation since it learns and combines multi-scale

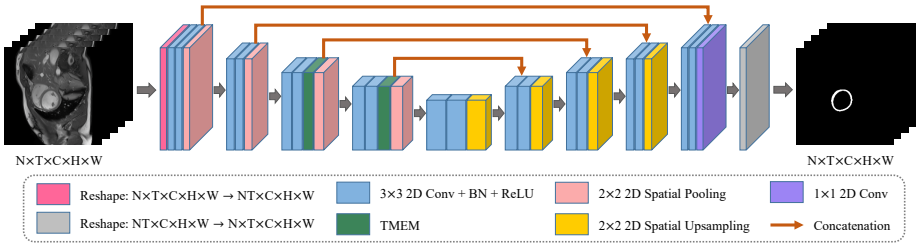


Fig. 1. Overview of the proposed method. The input is a volumetric sequence of cardiac images, and the output is the corresponding LV segmentation results. The overall structure of our method is based on (but not limited to) a 2D U-Net, where we insert two Temporal Motion Encoding Modules (TMEM) to effectively exploit the temporal information for cardiac MR image segmentation. N: the batch size; T: the number of frames in a sequence; H and W: the spatial size of feature maps; $C = 1$: the channel number for grayscale image.

features. This design has inspired many follow-up methods. However, this type of methods segment the slices individually without considering their spatial and temporal correlations. To address this issue, the 3D U-Net [2] extends the 2D U-Net by replacing the 2D convolutions with 3D ones in order to capture long-range dependencies between different slices. While achieving improved accuracy, 3D U-Net inevitably increases the computational cost and tends to cause overfitting. Such weaknesses can be alleviated by the recurrent U-Net [6], which employs ConvGRU [1] to connect the slices but still suffers from computational inefficiency. Recent works [7, 11] have sought to exploit optical flow for capturing cardiac dynamic features and enforcing temporal coherence. However, the extraction of optical flow is non-trivial, expensive and prone to significant errors, making these methods often inaccurate and difficult to deploy in real applications. Other attempts at improving U-Nets include adopting a hybrid solution [12] or integrating the attention mechanism into the network for feature refinement [5]. However, these methods are difficult to train, and fail to explicitly and efficiently model temporal relationships.

To address the above limitations, we propose a novel method for the spatio-temporal segmentation of cardiac MR image sequences. Our method is based on a 2D U-Net, and aggregates the temporal features with only 1D and 2D convolutions, thereby eliminating the heavy computation and massive parameters of 3D and recurrent convolutions. Specifically, we construct a Temporal Aggregation Module (TAM) to capture the inter-slice temporal features. TAM reformulates the input feature map as a 1D signal, and utilizes 1D convolution for temporal feature learning with small extra computation and parameter overhead. In addition, we introduce a Motion Encoding Module (MEM), which explicitly models the cardiac motion features using 2D convolutions without relying on optical flow. By integrating such dynamic information into a 2D U-Net, MEM is able to guide and regulate the segmentation. Finally, we integrate these two modules in a Temporal Motion Encoding Module (TMEM), which feeds the network with

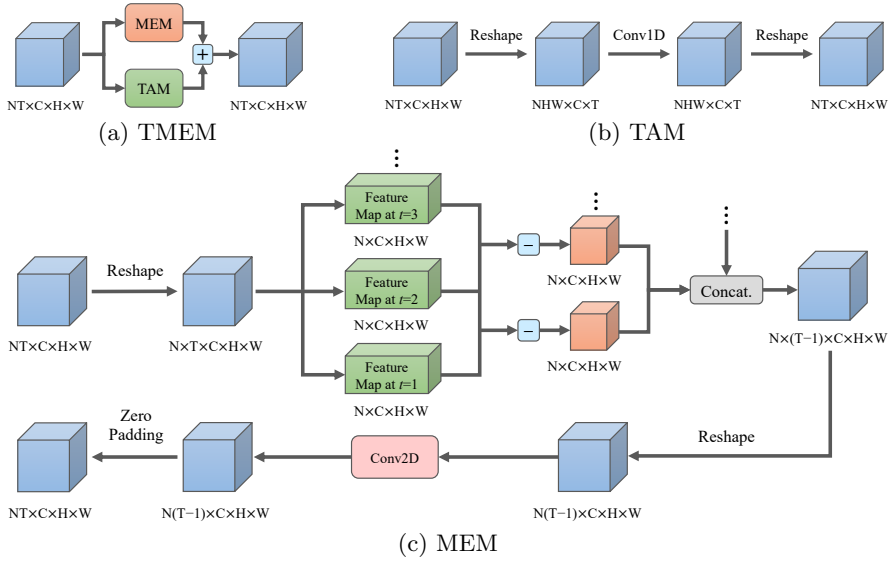


Fig. 2. Illustration of the different modules. (a) Temporal Motion Encoding Module (TMEM). (b) Temporal Aggregation Module (TAM). (c) Motion Encoding Module (MEM). These modules are computationally efficient, and enable to significantly improve the performance of base network with small extra overhead. "+" and "-" denote element-wise addition and subtraction, respectively. The Conv2D block consists of one 3×3 2D convolution layer, followed by BatchNorm and ReLU.

complementary temporal and motion information while preserving the simplicity and efficiency of 2D U-Net (see Fig. 1 and Fig. 2). Experimental results on two cardiac MR image datasets demonstrate the effectiveness and superiority of the proposed method compared to the state-of-the-arts.

2 Method

The overall architecture of our method is illustrated in Fig. 1, where we integrate two Temporal Motion Encoding Modules into the classic 2D U-Net. Below we give the details of our module design and the corresponding motivation.

2.1 Temporal Aggregation Module

The Temporal Aggregation Module (TAM) aims to extract the temporal features with very limited extra overhead. As shown in Fig. 2(b), given a 2D feature map $\mathbf{F} \in \mathbb{R}^{NT \times C \times H \times W}$, we first reshape it into a 1D signal $\mathbf{F}' \in \mathbb{R}^{NHW \times C \times T}$. Then, we apply 1D convolution to $\mathbf{F}'$ along the dimension of T to aggregate temporal features. This design has the benefit that the temporal information is propagated among different slices with only 1D convolution, which requires a

very small number of parameters and computational cost. In our implementation of TAM, we use a 1D convolution with kernel size 3.

2.2 Motion Encoding Module

Existing works, such as [7, 11], have indicated that the optical-flow motion information is able to regulate the network and thereby significantly improve the segmentation performance. However, the computation or learning of optical flow in their methods is non-trivial, computationally expensive and error-prone, which hinders their practical applications. To address this issue, we design a Motion Encoding Module (MEM), which aims to capture the motion features efficiently rather than recover the exact motion patterns as in optical flow. To be specific, the motion information from MEM is at the feature level, and can be computed efficiently and used to improve the segmentation.

Specifically, given a feature map $\mathbf{F} \in \mathbb{R}^{NT\times C\times H\times W}$, we first reshape it to expose the temporal dimension, obtaining $\mathbf{F}' \in \mathbb{R}^{N\times T\times C\times H\times W}$. Then we split $\mathbf{F}'$ into a set of feature maps $\mathbf{F}'_1, \ldots, \mathbf{F}'_T$, where $\mathbf{F}'_t \in \mathbb{R}^{N\times C\times H\times W}, t \in [1, T]$. Afterwards, the motion information is extracted from every two consecutive feature maps $\mathbf{F}'_t$ and $\mathbf{F}'_{t+1}$. Formally,

$$\widetilde{\mathbf{F}}_t = f(\mathbf{F}'_t - \mathbf{F}'_{t+1}), \quad t \in [1, T-1], \tag{1}$$

where $\widetilde{\mathbf{F}}_t \in \mathbb{R}^{N\times C\times H\times W}$ is the captured motion information, and f denotes a nonlinear function, which in our case is a 2D convolution followed by BatchNorm and ReLU. Finally, all the generated motion features $\widetilde{\mathbf{F}}_t$ are stacked along the temporal dimension, providing the feature map $\widetilde{\mathbf{F}} \in \mathbb{R}^{N\times(T-1)\times C\times H\times W}$. To make the size of $\widetilde{\mathbf{F}}$ consistent with the input feature $\mathbf{F}$, we pad $\widetilde{\mathbf{F}}$ with zeros at the last time step, and reshape it into an output feature map $\widehat{\mathbf{F}} \in \mathbb{R}^{NT\times C\times H\times W}$ (see Fig. 2(c)).

As is illustrated, the proposed MEM is simple and only relies on 2D convolution, and thus it is more efficient than the 3D and recurrent counterparts. In the experiments, we will show that MEM is able to largely improve the segmentation performance of a basic 2D U-Net.

2.3 Temporal Motion Encoding Module

TAM and MEM extract the temporal features from two different perspectives. To combine their strengths we design a Temporal Motion Encoding Module (TMEM), which consists TAM and MEM (see Fig. 2(a)). TMEM is able to fuse the temporal and motion features together and can be integrated into any layer of a 2D U-Net. In practice, we empirically observe that placing TMEM within the third and fourth Conv2D blocks of the encoder gives the best results. This observation is in accordance with the findings of [10], which suggest temporal representation learning on high-level semantic features is more useful.

3 Experiments and Results

We evaluate the proposed method on two cardiac MR image datasets. (1) DYS, a dataset which contains 24 subjects, of which the patients are with heart failure due to dyssynchrony. The number of phases is 25 for each cardiac cycle. In total, there are around 4000 2D short-axis (SAX) slices. The LV myocardium contours of these SAX images over different spatial locations and cardiac phases are manually annotated based on consensus of three medical experts. The sizes of images vary between 224×204 and 240×198 pixels, and their in-plane resolutions vary from $1.17mm$ to $1.43mm$. We use 3-fold cross validation in our experiments, and make sure both the training and test sets contain the normal subjects and patients. (2) CAP, a publicly available dataset consisting of steady-state free precession (SSFP) cine MR images from Cardiac Atlas database [3, 4, 9]. CAP involves 100 patients with coronary artery disease and prior myocardial infarction. The ground-truth myocardium annotations are generated by various raters with consensus. There exists large variability within this dataset: the data are generated from different MRI scanner systems, the image size varies from 138×192 to 512×512 and the cardiac phases range from 18 to 35. These factors make CAP more challenging than DYS. In our experiments, we perform cross validation with 3 different partitions of the dataset. In each particular partition, we select 70 subjects for training, 15 for validation and 15 for testing.

During training, for both datasets, we crop the regions around the LV to generate training images of size 144×144. Data augmentation, including random flip and rotation, is adopted to improve the model robustness. To train the models, we use cross entropy loss and optimize the network parameters with Adam optimizer. We set the learning rate to 0.0005, and decay it by 0.5 after every 15 training epochs. The batch size is 8 (i.e., 8 cardiac sequences), each of which is padded/subsampled to contain 32 slices. The weight decay is 0.0001, and the number of training epochs is chosen to be 75 to ensure convergence. For all the 2D convolutions of our method, we set their kernel sizes to 3×3. The training of our model takes around $0.5 \sim 1.5$ hours on a single NVIDIA RTX 4000 GPU, and the inference takes about 0.01s for a sequence of slices.

We compare our method with several representative works, including 2D U-Net [8], 3D U-Net [2], recurrent U-Net (RFCN) [6], and Attention U-Net [5]. In particular, the 3D U-Net was originally developed to capture the spatial relationships among the slices of 3D volumetric images (i.e., stacks of images). Contrary to its original application, here we apply the 3D U-Net to a sequence of SAX images from a cardiac cycle, and set its channel numbers equal to its 2D counterpart. Similarly, RFCN was designed for processing the slices of a single 3D volumetric cardiac image, and here we apply its recurrent unit to the temporal dimension. The Attention U-Net aims to improve the classic U-Net, and generates attention maps from higher-level features to help the network focus on important regions. Apart from the above methods, we also build another U-Net variant which is inspired by [10]. Specifically, we replace the TMEM in our method with a separable 3D convolution, which decomposes the traditional $3 \times 3 \times 3$ 3D convolution into two separate ones: a 3D convolution with kernel

Table 1. Evaluation of segmentation accuracy for different methods in terms of Dice and Jaccard metrics, as well as Hausdorff distance (HD) in pixels. We report the mean and standard deviation over different folds.

Dataset	Method	Dice	Jaccard	HD
DYS	2D U-Net	0.7854 ± 0.0384	0.6633 ± 0.0465	4.6204 ± 2.8586
	3D U-Net	0.7566 ± 0.0143	0.6229 ± 0.0175	10.159 ± 3.4157
	Att. U-Net	0.7984 ± 0.0269	0.6801 ± 0.0303	5.9199 ± 1.3563
	RFCN	0.7936 ± 0.0316	0.6721 ± 0.0391	4.5095 ± 1.3524
	S3D U-Net	0.7912 ± 0.0348	0.6719 ± 0.0393	5.1291 ± 1.6801
	Ours + TAM	0.8101 ± 0.0254	0.6944 ± 0.0288	3.8381 ± 0.7819
	Ours + MEM	0.8076 ± 0.0203	0.6922 ± 0.0197	3.9163 ± 1.1508
	Ours + TMEM	$\mathbf{0.8204 \pm 0.0302}$	$\mathbf{0.7085 \pm 0.0363}$	$\mathbf{3.4114 \pm 0.7277}$
CAP	2D U-Net	0.7158 ± 0.0243	0.6234 ± 0.0271	4.4257 ± 0.4978
	3D U-Net	0.7434 ± 0.0084	0.6501 ± 0.0051	4.2577 ± 0.3889
	Att. U-Net	0.7341 ± 0.0124	0.6445 ± 0.0127	4.6684 ± 0.1695
	RFCN	0.7172 ± 0.0297	0.6259 ± 0.0278	4.6394 ± 0.3271
	S3D U-Net	0.7191 ± 0.0118	0.6276 ± 0.0101	5.3890 ± 0.3459
	Ours + TAM	0.7653 ± 0.0175	0.6766 ± 0.0188	4.3039 ± 0.4061
	Ours + MEM	0.7681 ± 0.0213	0.6814 ± 0.0206	4.5178 ± 0.5474
	Ours + TMEM	$\mathbf{0.7766 \pm 0.0087}$	$\mathbf{0.6912 \pm 0.0066}$	$\mathbf{3.8977 \pm 0.2190}$

Table 2. The model complexities of different methods. FLOPs are calculated over a sequence of 32 images, with size 144×144.

Method	#Parameter	FLOPs	Method	#Parameter	FLOPs
2D U-net	7.9M	143G	Att. U-Net	8.5M	150G
3D U-Net	23.5M	183G	S3D U-Net	8.8M	159G
RFCN	22.0M	179G	Ours + TAM	8.1M	146G
Ours + MEM	8.6M	154G	Ours + TMEM	8.8M	158G

size $1 \times 3 \times 3$ followed by another one with kernel size $3 \times 1 \times 1$. We term this model as S3D U-Net, which only requires a small extra parameter and computation overhead while being able to capture the temporal information. Finally, to validate the effectiveness of our two modules, we also conduct ablation studies. In particular, we remove MEM from our model, obtaining a network with TAM only (i.e., Ours + TAM in Table 1). Similarly, we remove TAM and replace it with an identity mapping, leading to a model with MEM only (i.e., Ours + MEM in Table 1).

Table 1 lists the segmentation results for different methods. As is shown, both the TAM and MEM are able to significantly boost the performance of vanilla 2D U-Net. This demonstrates the importance of leveraging temporal information in cardiac image sequence segmentation, as well as the effectiveness of the proposed modules on exploiting temporal features. In addition, when combining TAM and MEM into a single module, we observe a further improved segmentation accuracy, which indicates that the temporal and motion features are complementary to each other. From Table 1 it can also be observed that our method outperforms

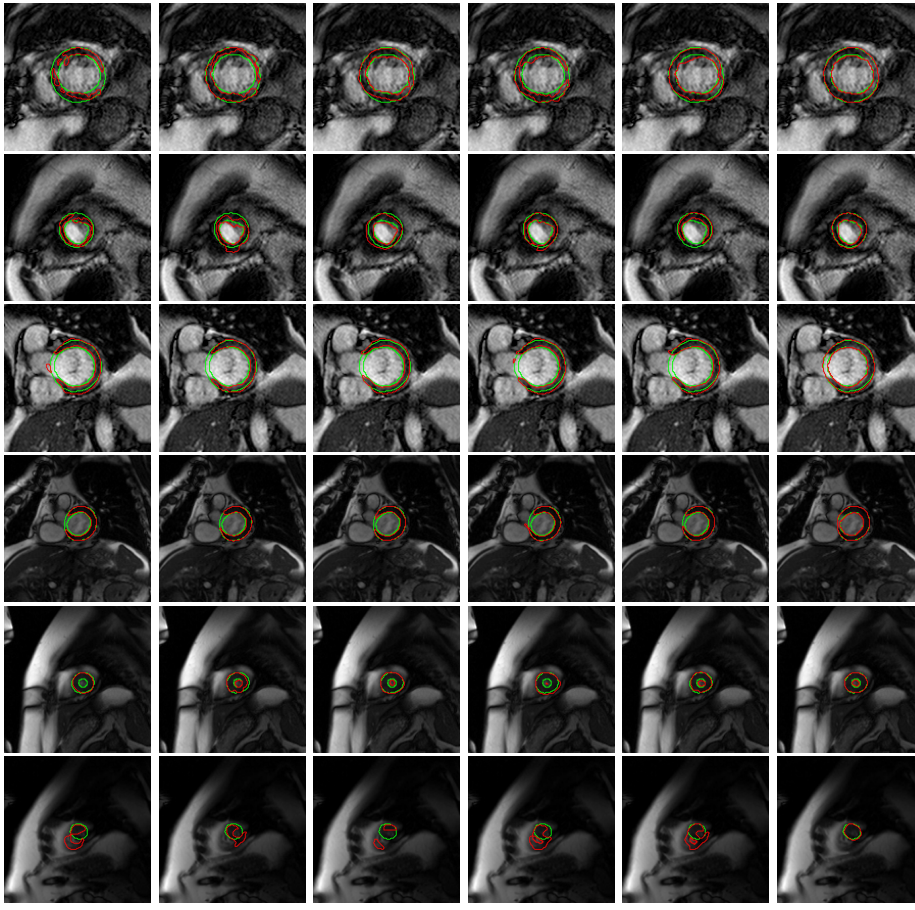
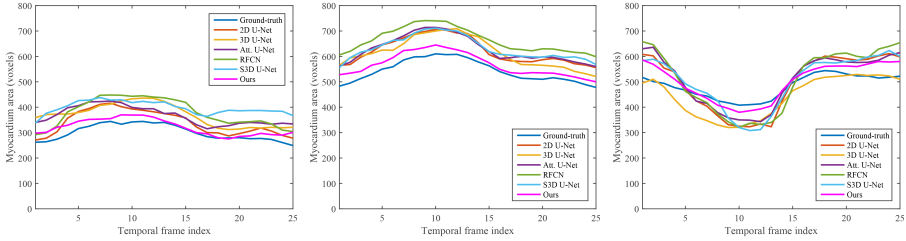


Fig. 3. Examples of segmented left ventricle walls from different methods. The images are from the CAP dataset, and from different LV locations: middle (row 1-2), base (row 3-4) and apex (row 5-6). These images show the cases of myocardial infarction; our method overall achieves the best performance. Green contours represent the ground truth and red contours are the model predictions. "Ours" refers to the model using TMEM. (Zoom in for best view.)

RFCN, 3D U-Net and S3D U-Net, even when only one of the proposed modules is employed. This validates the advantages of our temporal feature encoding over the recurrent 2D and vanilla/separable 3D convolutions. In Table 2 we also report the model complexities of different methods. It can be observed that our modules only introduce a small extra parameter and computation overhead while bringing a clear performance gain.

Fig. 3 shows several segmentation results for different methods. We can observe that our method is able to delineate the ventricular walls accurately, espe-



(a) Apex (b) Middle (c) Base

Fig. 4. Representative examples of LV myocardium area over a cardiac cycle for different methods, at the apex, middle and base, respectively. The data sample is from the CAP dataset, and shows a case of myocardial infarction.

cially at the base and the apex which are challenging (see row 3-6). Moreover, the proposed method is able to generate smoother outcomes while preserving the topological shape properties of the myocardium walls (i.e., a closed loop). In contrast, the original 2D U-Net fails to accurately localize the boundary at the base and apex, and leads to disconnected segmented shapes. This demonstrates the effectiveness of our method on leveraging temporal and motion information for the segmentation of cardiac MR image sequences.

In Fig. 4, we plot the myocardium area over a cardiac cycle for different methods. We observe that, compared to the baselines, the results by our method are smoother and closer to the ground-truth. In particular, our method largely improves the 2D U-Net, thanks to the explicit modeling of temporal information.

4 Conclusions

In this work, we proposed a new method for the segmentation of cardiac MR image sequences, based on the use of a 2D U-Net. The key elements of our method are two new modules, which are able to leverage the temporal and motion information volumetrically in cardiac image sequences. The proposed modules work collaboratively and enable us to improve the feature learning of the base network in a computationally efficient manner. Experimental results on two cardiac MR image datasets demonstrate the effectiveness of our method.

References

1. Ballas, N., Yao, L., Pal, C., Courville, A.: Delving deeper into convolutional networks for learning video representations. In: ICLR (2016)
2. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: MICCAI. pp. 424–432. Springer (2016)
3. Fonseca, C.G., Backhaus, M., Bluemke, D.A., Britten, R.D., Chung, J.D., Cowan, B.R., Dinov, I.D., Finn, J.P., Hunter, P.J., Kadish, A.H., et al.: The cardiac atlas

project—an imaging database for computational modeling and statistical atlases of the heart. Bioinformatics **27**(16), 2288–2295 (2011)
4. Kadish, A.H., Bello, D., Finn, J.P., Bonow, R.O., Schaechter, A., Subacius, H., Albert, C., Daubert, J.P., Fonseca, C.G., Goldberger, J.J.: Rationale and design for the defibrillators to reduce risk by magnetic resonance imaging evaluation (determine) trial. Journal of cardiovascular electrophysiology **20**(9), 982–987 (2009)
5. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al.: Attention u-net: Learning where to look for the pancreas. In: MIDL (2018)
6. Poudel, R.P., Lamata, P., Montana, G.: Recurrent fully convolutional neural networks for multi-slice MRI cardiac segmentation. In: Reconstruction, segmentation, and analysis of medical images, pp. 83–94. Springer (2016)
7. Qin, C., Bai, W., Schlemper, J., Petersen, S.E., Piechnik, S.K., Neubauer, S., Rueckert, D.: Joint learning of motion estimation and segmentation for cardiac MR image sequences. In: MICCAI. pp. 472–480. Springer (2018)
8. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241. Springer (2015)
9. Suinesiaputra, A., Cowan, B.R., Al-Agamy, A.O., Elattar, M.A., Ayache, N., Fahmy, A.S., Khalifa, A.M., Medrano-Gracia, P., Jolly, M.P., Kadish, A.H., et al.: A collaborative resource to build consensus for automated left ventricular segmentation of cardiac mr images. Medical image analysis **18**(1), 50–62 (2014)
10. Xie, S., Sun, C., Huang, J., Tu, Z., Murphy, K.: Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In: ECCV. pp. 305–321 (2018)
11. Yan, W., Wang, Y., Li, Z., Van Der Geest, R.J., Tao, Q.: Left ventricle segmentation via optical-flow-net from short-axis cine MRI: preserving the temporal coherence of cardiac motion. In: MICCAI. pp. 613–621. Springer (2018)
12. Yang, D., Huang, Q., Axel, L., Metaxas, D.: Multi-component deformable models coupled with 2d-3d u-net for automated probabilistic segmentation of cardiac walls and blood. In: ISBI. pp. 479–483 (2018)