Joint semantic feature and optical flow learning for automatic echocardiography segmentation

Juan Lyu[1[0000-0002-1366-1807]], Jinpeng Meng[2[0009-0000-5312-8655]], Yu Zhang[1*[0000-0002-8954-6741]], Sai Ho Ling[3[0000-0003-0849-5098]], Lin Sun[1[0000-0003-4917-7651]]

[1] College of Artificial Intelligence, Tianjin University of Science and Technology, Tianjin 300457, China
Lvjuan@tust.edu.cn, Zhangyuai@tust.edu.cn, Sunlin@tust.edu.cn

[2] College of Light Industry Science and Engineering, Tianjin University of Science and Technology, Tianjin 300457, China
22062212@mail.tust.edu.cn

[3] School of Electrical and Data Engineering, Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, NSW 2007, Australia
Steve.Ling@uts.edu.au

Abstract. The left ventricle ejection fraction is an important index for assessing cardiac function and diagnosing cardiac diseases. At present, EchoNet-Dynamic dataset is the unique large-scale resource for studying ejection fraction estimation by echocardiography. Through segmentation of the end-systolic and end-diastolic frames, the ejection fraction can be calculated based on the volumes at these phases. However, existing segmentation methods either mostly focus on single-frame segmentation and rarely consider information across consecutive frames, or they fail to effectively exploit temporal information between consecutive frames, resulting in suboptimal segmentation performance. In our study, we constructed a dual-branch spatial-temporal feature extraction model for achieving echocardiogram video segmentation. One branch was dedicated to extracting semantic features of frames under supervision, while the other branch learned the optical flows between frames in an unsupervised manner. Subsequently, we jointly trained these two branches using a temporal consistency mechanism to acquire spatial-temporal features of the frames. This approach enhances both video segmentation performance and the consistency of transition frame segmentation. Experimental results demonstrate that our proposed model achieves promising segmentation performance compared to existing methods.

Keywords: Echocardiography Segmentation, Optical Flow, Joint Learning.

1 Introduction

Cardiovascular diseases are the leading cause of death worldwide, accounting for 32% of the total global deaths, of which heart attack and stroke represent 85% of these deaths. It is recommended by World Health Organization (WHO) that early diagnosing is crucial for cardiovascular diseases. For evaluating heart function and structure, echocardiography is a commonly utilized tool in any stage of clinical practice [1]. At present, deep learning has been the most popular way of echocardiography segmentation task and achieved much better performance. Paper [2, 3] utilized U-Net-based networks to segment the ES and ED frames. Li et al. proposed a multi-level and multi-scale dense pyramid and deep supervision network (DPSN) for segmentation of key frames in multi-chamber views [4]. Other approaches [5, 6] integrated convolutional neural network (CNN) models with transformer modules to utilize image patches for segmentation. Some researchers have also incorporated attention techniques to enhance feature fusion effectiveness for segmentation[6-8]. However, the above single-image segmentation methods typically overlook the temporal information and inter-frame correlations between video frames, resulting in challenges in accurately delineating the left ventricular region, particularly in intermediate transition frames.

Recently, more studies started to focus on the echocardiographic video segmentation, which located the ES and ED frames based on the volumes obtained by the segmentation of all frames. To introduce temporal information, some of the methods adopted 3D structures to extract the semantic and temporal features at the same time. For example, Wei et al. proposed a co-learning network that trains both at the appearance level and the shape level based on 3D U-Net [9, 22]. Chen et al. proposed a 3D U-Net for echocardiography video segmentation by learning the ED and ES segmentation and motion tracking between the frames at the same time [10]. However, the 3D-based networks cannot be used in single image cases, which has limitations in clinical practice. Other approaches employed the 2D plus time (2D + t) architecture to discover spatial-temporal information, which take videos or image sequences as inputs. Li et al. proposed a multi-view echocardiographic video segmentation network based on long-short term memory (LSTM), named MV-RAN [11]. Although the MV-RAN can model the temporal consistency, the LSTM structure is time-consuming and causes the end frames of the video to perform worse than the beginnings due to the errors accumulated. Sirhani et al. proposed a EchoRCNN model based on the mask region-based CNN (Mask RCNN) [12]. However, the ground truth mask of the first frame of the video should be delineated, which increases the cost of clinical application. Moreover, the proposed EchoRCNN was validated on a small dataset with only 750 videos. Painchaud et al. proposed an enforced temporal consistency post-processing approach to achieve echocardiographic video segmentation [13]. However, its performance improvement is limited. Wu et al. proposed an adaptive spatiotemporal semantic calibration (ASSC) module to utilize the spatio-temporal information between consecutive frames and to overcome the drawback that the optical-flow-based models are sensitive to speckle noise [14]. However, the ASSC module used a series of transformations and imported several learnable transformation metrics for both coordinate warping calibration and channel-wise feature weighting calibration, which made the model more complex and difficult to learn these metrics.

In this research, we introduced a novel dual-branch spatial-temporal joint learning network for echocardiographic video segmentation. The network consists of a 2D image segmentation branch to learn the spatial features of the inputs and to achieve the frames segmentation, and an optical flow learning branch to extract the optical flow between every two frames. Based on the optical flow learned from two consecutive frames, we jointly learned spatial and temporal information using a temporal consistency module between the warped segmentation prediction and the real segmentation prediction at t time. The contributions of this paper are as follows.

- We developed a dual-branch network which consists of a supervised semantic segmentation branch, and an unsupervised optical flow learning branch to learn the consistency between the consecutive frames.
- We jointly trained the two branches using the temporal consistency technique to learn the spatial-temporal features of the videos.
- The proposed model achieved a promising segmentation performance on the EchoNet-Dynamic dataset and demonstrated higher consistency in transition frames than other approaches.

2 Methods

In this work, we presented a dual-branch echocardiographic video segmentation approach that uses video clips as inputs. As illustrated in Fig. 1, the proposed network consists of two branches. The segmentation branch was employed to segment the left ventricle area in each frame. The optical flow branch was used to learn the optical flow changes and temporal information between frame pairs. Finally, we jointly trained two branches by the proposed temporal consistency mechanism.

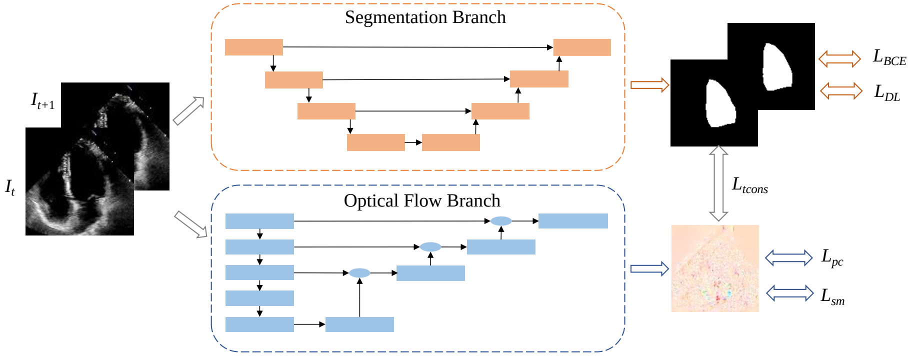

Fig. 1. The architecture of the proposed echocardiography segmentation network. The structure of each branch is presented in corresponding box roughly.

2.1 Overview of Framework Workflow

The architecture of the proposed model is a spatial and temporal combination structure, composed of two branches: the segmentation branch and the optical flow branch. The videos in the EchoNet-Dynamic dataset are typically large, with an average duration of more than 176 frames, while only two frames in each video are labeled. When training the frames in pairs, only two frames can be used to update the segmentation branch, while all frame pairs are used to update the optical flow branch..

In this paper, in the training stage, we set two clips for each video, the ES frame and its former and later two frames as clip one, the ED frame and its former and later two frames as clip two. They are defined as $c1$: $\{I_{ES-1}, I_{ES}, I_{ES+1}\}$ and $c2$: $\{I_{ED-1}, I_{ED}, I_{ED+1}\}$. All clips were used in the training in pairs to learn the semantic segmentation and optical flow parallelly according to the model shown in Fig. 1. In the testing stage, we tested all the frames of each video and output their predicted left ventricle masks only using the segmentation branch.

2.2 Segmentation Learning

For the segmentation branch, we adopted a 2D image segmentation network to learn the spatial semantic features of the input echocardiography. The main target of this branch is to distinguish between the region of interest (left ventricle) and the background. Therefore, in this branch, we adopted the baseline model U-Net to focus on spatial semantic feature extraction, more details can be found in paper [15].

As shown in Fig. 2, the input images are trained in pairs between two consecutive frames, denoted as I_t and I_{t+1}. We represented the segmentation branch as $S_g(x)$, where g is its corresponding parameter, and simply referred to it as the S branch for convenience. The corresponding outputs of two input pairs are $S_g(I_t)$ and $S_g(I_{t+1})$, respectively. The S branch was trained using two common semantic segmentation loss functions: binary cross-entropy (BCE) loss and dice loss (DL), which are defined as

$$L_{BCE} = -y \log \hat{y} - (1-y)\log(1-\hat{y}) , \tag{1}$$

where y and $\hat{y}$ denote semantic region label and the predicted result, respectively.

$$L_{Dice} = 1 - \frac{2|Y \cap G|}{|Y| + |G|} , \tag{2}$$

where we set the predicted segmentation results as Y and its corresponding label as G; the numerator denotes the twice of the overlap area of two sets Y and G, the denominator is the sum of elements in the two sets.

The total loss function of the S branch is defined as

$$L_S = L_{BCE} + L_{Dice} . \tag{3}$$

Notably, the segmentation learning was supervised, with human experts annotating the masks. That is, the segmentation branch can only output their predicted masks for frames without mask labels; they cannot be used to update the weights of the network.

2.3 Optical Flow Learning

For the optical flow branch, we employed a specialized network to learn temporal information between two adjacent frames through the optical flow. Compared to region-based networks, it is more suitable to use a pixel-level algorithm to discover the pixel-scale movement between two consecutive frames. In particular, most of the brightness changes occur at the edge of the heart chambers, which can also help to distinguish the edge from the background.

In this section, we designed a modified FlowNet based on FlowNetSimple [16]. Fig. 2 illustrates the architecture of the modified FlowNet, denoted as mFlowNet. The blue component is derived from the original FlowNetSimple, which we customized by importing part of layers. The green section represents our modifications, in which we added more up-sampled layers to ensure that the outputs are of the same size as the inputs. The reason is that we hope to use deconvolutions to learn the up-sampling process, instead of the interpolation during the warping computation. The corresponding hyperparameters for each operation are provided below them in Fig. 2, where *f* denotes the number of features, *k* denotes the kernel size of the convolution, *s* denotes the step size, *p* denotes the padding size. The number of features of the deconvolution in refine operation is specified below the Refine block. “Up flow” represents the up-sampled operation to predict flow. In mFlowNet, we also adopted the encoder and decoder structures to learn the optical flow between every two frames. In detail, it contains five normal convolution and down-sampling blocks in the encoder. For the decoder, we introduced two additional up-sampling layers and one more feature fusion layer to ensure that the output size matches that of the input.

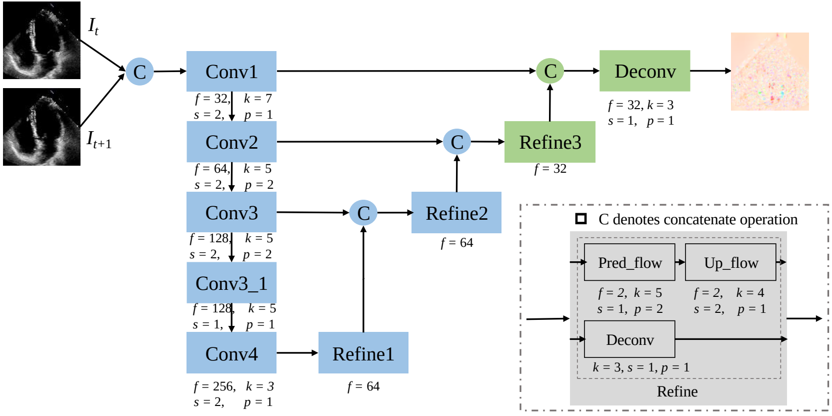

Fig. 2. The architecture of the mFlowNet. The blue rectangles represent the original FlowNet blocks, while the green rectangles represent the modified parts by this work.

We represented the optical flow branch as $O_p(x)$, where *p* is its corresponding parameter, and simply named it as O branch. The inputs of the O branch are still in pairs, I_t and I_{t+1}, which are the same as the inputs of the S branch. Two frames of inputs were concatenated in pairs at the channel level, forming a 6-channel input. The output of the mFlowNet is the optical flow between the two input frames, presented as $M_{t\to t+1}$. The mFlowNet was trained in an unsupervised manner and its update was depended on the basic characteristics of optical flow, photometric consistency and motion smoothness.

Photometric consistency loss [16, 18] is to constrain a frame and the warped image from its adjacent frame, which is defined as

$$L_{pc} = \alpha \frac{1 - SSIM(I - I_w)}{2} + (1 - \alpha) \| I - I_w \|_1 , \tag{4}$$

where I_w is the warped image, $SSIM$ is the structural similarity index and α is set to 0.85 accordingly [18]. The purpose of motion smoothness is intended to eliminate erroneous predictions while preserving crisp details, which is defined as

$$L_{sm} = \sum_{x,y} | \nabla M(x, y) | \cdot (e^{-|\nabla I(x,y)|}) , \tag{5}$$

where ∇ is the vector differential operator, $|\cdot|$ denotes element-wise absolute value. The total loss function for O branch is presented as

$$L_O = \lambda_1 L_{pc} + \lambda_2 L_{sm} , \tag{6}$$

where λ_1 and λ_2 is the corresponding weights of two losses, respectively.

2.4 Cooperation Mechanism and Joint Learning

For the above two branches, the S branch is to learn the spatial semantic features, and the O branch is to discover the temporal features between the frames. We utilized temporal consistency constraints to fuse the learned features to further improve the segmentation performance. We adopted the temporal consistency module in [19]. They defined the temporal consistency constraint as the function of the encoder output features at time *t* and the warped features from time *t* + 1. However, in this paper, the temporal consistency constraint is defined as a function of the segmentation output at time *t* and the warped output from time *t* + 1 using the learnt optical flow from the O branch. The rationale behind this choice is that the edges between the left ventricle and the background tend to be blurred in ultrasound imaging. Therefore, the temporal consistency module that only works on the segmentation output can help filter out the background and noise from non-left ventricle regions. Since the segmentation output is binary, that is, the pixel values of the segmented background are all zero, only the segmented left ventricle region is used for the optical flow warping computation.

Given a pair of input frames I_t and I_{t+1}, we got their semantic segmentation results from branch S, Y_t and Y_{t+1}, respectively, and obtained their predicted optical flow from branch O, $M_{t \to t+1}$. Then we warped Y_{t+1} to $Y_t^{'}$ by optical flow $M_{t \to t+1}$, which is calculated by

$$Y_t^{'} = \mathrm{Warp}(Y_{t+1}, M_{t \to t+1}) , \tag{7}$$

where we also used differentiable bilinear interpolation for warping. Since our dataset does not have the occluded issue, the temporal consistency loss is defined as

$$L_{tcons} = \sum_{x,y} \| Y^{'xy} - Y^{xy} \| . \tag{8}$$

In this way, we introduced temporal features into spatial space through optical flow and warping. Consequently, we are able to use the temporal O branch to extract features from the unlabeled frames and then enhance the semantic segmentation result through warping. Two branches work together in an end-to-end manner to achieve the video segmentation and improve the performance of the model.

The total loss function of the proposed model is

$$L = L_S + L_O + \lambda_3 L_{tcons} = L_{BCE} + L_{Dice} + \lambda_1 L_{pc} + \lambda_2 L_{sm} + \lambda_3 L_{tcons} \ , \tag{9}$$

where the weights of L_S and L_O are set to 1, the weights of L_{tcons} is λ_3 .

3 Materials

3.1 Data

EchoNet-Dynamic is a large-scale, publicly available echocardiography video dataset for cardiac function assessment that we employed in this paper. The EchoNet-Dynamic dataset contains 10,030 echocardiographic videos recorded independently by 10,030 people. For each video, the video length, the positions (time points), masks and volumes of ES and ED frames, and the correspondingly calculated EF are provided. The size of all the frames in the dataset is 112 × 112. All the annotations are supplied by experienced experts.

3.2 Implementation Details

The experiments were implemented using the Pytorch library version 1.6.0. The training and testing were done on a machine with an Intel Core i7-9700K CPU processor, 31.2 GB of memory, and a GeForce 2080 Ti 11GB GPU.

The dataset was divided into training, validation, and testing sets in the ratios of 75%, 12.5%, and 12.5%, respectively, which is the same as the setting of EchoNet-Dynamic [20]. For fair comparison with parts of other models, we also evaluate the proposed method following their training and testing ratio of 80%:20%. During the training stage, as mentioned previously, we utilized video clips to train the proposed model. Each clip generates four pairs of inputs for every video. In the testing stage, we tested all the frames in each video. We trained the model for 100 epochs with a batch size of one. We used the Adam optimizer to update the model weights with an initial learning rate of 1.6×10^{-5}. For the loss function, we experimentally set the λ_1 , λ_2 , and λ_3 to be 5, 0.2, and 0.4, respectively. In this work, we utilized Dice coefficient score and Hausdorff distance (HD) to evaluate the segmentation performance of the proposed model. Dice score is related to the dice loss and defined as

$$Dice(Y, G) = 1 - L_{Dice} \, . \tag{10}$$

HD is used to valuate the maximum distance between the prediction Y and ground truth G , HD is defined as

$$H(Y,G) = \max(h(Y,G), h(G,Y)) \,, \tag{11}$$

we take direct Hausdorff distance from Y to G as an example, it is presented as

$$h(Y,G) = \max_{y \in Y}(\max_{g \in G}(d(y,g))) \,, \tag{12}$$

where $d(y,g)$ denotes the Euclidean distance between y and g .

4 Experiment

First, we investigated the effectiveness of introducing temporal features into the spatial feature extraction network for left ventricle segmentation. Second, we evaluated relations between the performance of the proposed method with the number of samples in the training clips. Third, we validated the advancement of the proposed method by comparing it with the existing networks on the EchoNet-Dynamic dataset.

4.1 Evaluation of Introducing Optical Flow Branch

We evaluated the effectiveness of importing the optical flow branch by comparing it with the spatial semantic network, U-Net. The comparison results are shown in Table 1. It turns out that extracting both spatial and temporal features at the same time is better for video segmentation than extracting only spatial features. The temporal features contain the information between the adjacent frames, thereby the network can provide spatial-temporal information for neighboring frames in the videos.

Table 1. Evaluation of Introducing Optical Flow Branch

Structure	Dice score (%)
U-Net	88.76
U-Net + FlowNetSimple	92.50
U-Net + mFlowNet (this work)	**92.64**

4.2 Affection of Training Sample Numbers

In Table 2, we compared the results of training with different amounts of samples, including 6 (this work), 10, 18, and ES to ED frames.

Table 2. Comparison Results of Using Different Samples for Training

Num of Samples per Video	Dice score (%)
6 (this work)	**92.64**
10	92.44
18	92.12
ES to ED frames	89.44

It can be seen that the segmentation performance decreases as the number of samples increases. This suggests that when more unlabeled data is introduced, the segmentation heavily relies on accurate optical flow estimation. However, since the learning process of the optical flow is unsupervised, it may lead to an accumulation of errors during the warping phase if the training samples are too numerous, resulting in decreased segmentation accuracy. Therefore, the optimal training length for this project is 6 samples.

4.3 Comparison with Existing Methods

We compared the proposed model with existing approaches on the EchoNet-Dynamic dataset to validate its segmentation performance, as shown in Table 3. For the 2D ES and ED frames segmentation methods, we compared several algorithms, including the primary algorithm by Ouyang et al., the **EchoNet-Dynamic** method [20] and three recent models: TransBridge [5] (offering **TransBridge-B** and **TransBridge-L** variants), **PLANet** [7], and **Bi-DCNet** [8]. They were evaluated on the training, testing, and validation sets provided by the EchoNet-Dynamic dataset, with a ratio of 75:12.5:12.5, referred to as ratio-1 for convenience. For echocardiographic video segmentation algorithms, we compared two approaches: **Joint-Net** [21] and a recent network [14] named **BSSF-Net**. Training and testing sets were randomly selected from the EchoNet-Dynamic dataset in an 80:20 ratio, denoted as ratio-2. These methods employed 5-fold cross-validation for evaluation and did not include a separate validation set. For comparison, we evaluated our proposed model and the baseline EchoNet-Dynamic algorithm using both ratios.

Table 3. Comparison Result with Existing Methods

Methods	Year	Train/Val/Test:	75/12.5/12.5	Train/Val/Test:80/-/20	
		Dice Score(%)	HD(mm)	Dice Score (mean±STD)(%)	HD (mean±STD)(mm)
EchoNet-Dynamic	2020	91.97	2.32	93.79±0.22	2.27±0.47
Joint-net	2020	-	-	90.91±0.36	3.85±0.92
TransBridge-B	2021	91.39	4.41	-	-
TransBridge-L	2021	91.64	4.19	-	-
PLANet	2021	-	-	91.92±0.34	3.42±0.67
BSSF-Net	2022	-	-	92.87±0.16	2.93±0.72
Bi-DCNet	2023	92.25	-	-	-
Ours	2024	**92.64**	**2.23**	**96.99±0.12**	**1.76±0.47**

In Table 3, our proposed method achieves the best segmentation results in both data ratios. For ratio-1, we achieved a Dice score of 92.64%, which is 0.39% higher than Bi-DCNet. In ratio-2, the proposed model demonstrates outstanding performance with a mean Dice score of 96.99%, surpassing ESSF-Net by 4.12% and EchoNet-Dynamic algorithm by 3.2%. This suggests that our spatial-temporal joint learning model excels in identifying the blurred edges of the left ventricle. Additionally, it indicates that the joint learning of semantic features and optical flows better exploits spatial-temporal information compared to 2D image segmentation methods and strategies proposed by Joint-net and BSSF-Net.

The comparison results were then represented in two different ways. Firstly, we compared the segmentation results of expert-labeled ES and ED frames. As depicted in Fig. 3, it is evident from the orange boxes that the contours segmented by our proposed technique are closer to the labels than those segmented by the EchoNet-Dynamic algorithm, indicating that our method can more accurately segment the left ventricle borders.

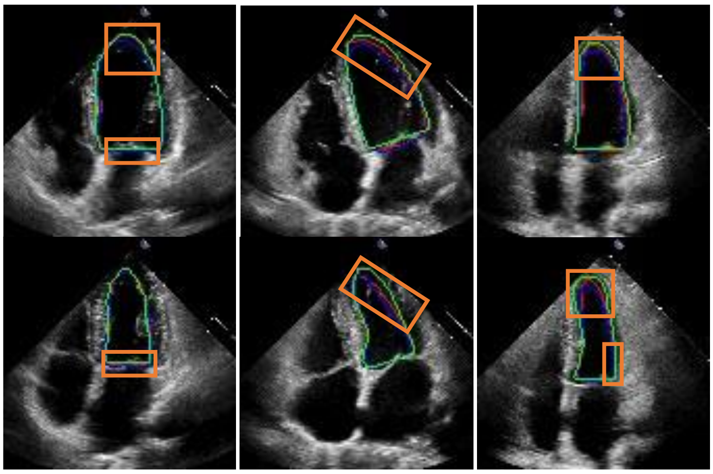

Fig. 3. Comparison results of ES and ED frames. Every column is an example of ES and ED frames in a video. The red circles are the results of this work, the blues are the results of the EchoNet-Dynamic algorithm, and the greens are the labels.

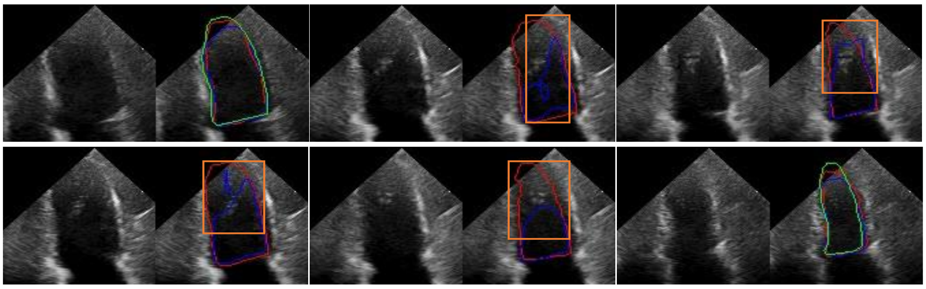

Fig. 4. Comparison results of unlabeled transition frames are depicted in pictures (a) and (b) for two separate videos, respectively. Each picture displays the original image on the left and the corresponding comparison visualization on the right.

Second, we exhibited the comparison results of unlabeled transition frames between this work and the EchoNet-Dynamic algorithm in Fig. 4. It can be seen that the EchoNet-Dynamic method was able to roughly segment the targets of ES and ED frames in the orange boxes. However, it is not able to distinguish targets in transition frames correctly, which is supposed to be affected by the imaging quality and noise. It indicates that the proposed method can not only more properly segment the ES and ED frames, but also more stably and reliably segment the transition frames in each video by learning the information between the key frames as well as the transition frames.

To summarize, the proposed method attained superior performance in echocardiography video segmentation by extracting the spatial-temporal properties of the frames. Compared to existing approaches, our method not only surpasses them in segmenting ES and ED frames but also demonstrates more consistent segmentation ability across other transition frames.

5 Conclusion

In this paper, we developed a novel echocardiography video segmentation network on the EchoNet-Dynamic dataset, which consists of a semantic features extraction branch and an optical flow learning branch. The two branches work together to combine the spatial and temporal information of the videos using a temporal consistency module to improve the performance of the left ventricle segmentation. The experimental results reveal that the proposed model achieves a promising performance compared with 2D ES and ED frames segmentation and echocardiographic video segmentation approaches, with a dice score of 92.46%. In the future, we will investigate more advanced temporal feature extraction strategies and the fuse mechanism to improve model segmentation performance.

References

1. K. T. Spencer, B. J. Kimura, C. E. Korcarz, P. A. Pellikka, P. S. Rahko, R. J. Siegel, Focused cardiac ultrasound: recommendations from the american society of echocardiography, Journal of the American Society of Echocardiography 26 (6) (2013) 567–581.
2. Y. Ali, F. Janabi-Sharifi, S. Beheshti, Echocardiographic image segmentation using deep res-u network, Biomedical Signal Processing and Control 64 (2021) 102248.
3. E. Puyol-Ant´ on, B. Ruijsink, B. S. Sidhu, J. Gould, B. Porter, M. K. Elliott, V. Mehta, H. Gu, M. Xochicale, A. Gomez, et al., Ai-enabled assessment of cardiac systolic and diastolic function from echocardiography, arXiv preprint arXiv:2203.11726 (2022).
4. M. Li, S. Dong, Z. Gao, C. Feng, H. Xiong, W. Zheng, D. Ghista, H. Zhang, V. H. C. de Albuquerque, Unified model for interpreting multi-view echocardiographic sequences without temporal information, Applied Soft Computing 88 (2020) 106049.
5. K. Deng, Y. Meng, D. Gao, J. Bridge, Y. Shen, G. Lip, Y. Zhao, Y. Zheng, Transbridge: A lightweight transformer for left ventricle segmentation in echocardiography, in: International Workshop on Advances in Simplifying Medical Ultrasound, Springer (2021) 63–72.
6. S. Shi, P. Alimu, P. Mahemuti, Q. Chen, H. Wu, The study of echocardiography of left-ventricle segmentation combining transformer and cnn, Available at SSRN 4184447 (2022).

7. F. Liu, K. Wang, D. Liu, X. Yang, J. Tian, Deep pyramid local attention neural network for cardiac structure segmentation in two-dimensional echocardiography, Medical Image Analysis 67 (2021) 101873.
8. Z. Ye, Y. J. Kumar, F. Song, G. Li, S. Zhang, Bi-dcnet: Bilateral network with dilated convolutions for left ventricle segmentation, Life 13 (4) (2023) 1040.
9. H. Wei, H. Cao, Y. Cao, Y. Zhou, W. Xue, D. Ni, S. Li, Temporal consistent segmentation of echocardiography with co-learning from appearance and shape, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer (2020) 623–632.
10. Y. Chen, X. Zhang, C. M. Haggerty, J. V. Stough, Assessing the generalizability of temporally coherent echocardiography video segmentation, in: Medical Imaging 2021: Image Processing, Vol. 11596, International Society for Optics and Photonics (2021) 463–469.
11. M. Li, C. Wang, H. Zhang, G. Yang, Mv-ran: Multiview recurrent aggregation network for echocardiographic sequences segmentation and full cardiac cycle analysis, Computers in biology and medicine 120 (2020) 103728.
12. N. Sirjani, S. Moradi, M. G. Oghli, A. Hosseinsabet, A. Alizadehasl, M. Yadollahi, I. Shiri, A. Shabanzadeh, Automatic cardiac evaluations using a deep video object segmentation network, Insights into Imaging 13 (1) (2022) 1–14.
13. N. Painchaud, N. Duchateau, O. Bernard, P.-M. Jodoin, Echocardiography segmentation with enforced temporal consistency, IEEE Transactions on Medical Imaging 41 (10) (2022) 2867–2878.
14. H. Wu, J. Liu, F. Xiao, Z. Wen, L. Cheng, J. Qin, Semi-supervised segmentation of echocardiography videos via noise-resilient spatiotemporal semantic calibration and fusion, Medical Image Analysis 78 (2022) 102397.
15. O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer (2015) 234–241.
16. A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, T. Brox, Flownet: Learning optical flow with convolutional networks, in: Proceedings of the IEEE international conference on computer vision, (2015) 2758–2766.
17. C. Godard, O. Mac Aodha, G. J. Brostow, Unsupervised monocular depth estimation with left-right consistency, in: Proceedings of the IEEE conference on computer vision and pattern recognition, (2017) 270– 279.
18. Z. Yin, J. Shi, Geonet: Unsupervised learning of dense depth, optical flow and camera pose, in: Proceedings of the IEEE conference on computer vision and pattern recognition, (2018) 1983–1992.
19. M. Ding, Z. Wang, B. Zhou, J. Shi, Z. Lu, P. Luo, Every frame counts: joint learning of video segmentation and optical flow, in: Proceedings of the AAAI Conference on Artificial Intelligence, 34 (2020) 10713–10720.
20. D. Ouyang, B. He, A. Ghorbani, N. Yuan, J. Ebinger, C. P. Langlotz, P. A. Heidenreich, R. A. Harrington, D. H. Liang, E. A. Ashley, et al., Video-based ai for beat-to-beat assessment of cardiac function, Nature 580 (7802) (2020) 252–256.
21. K. Ta, S. S. Ahn, J. C. Stendahl, A. J. Sinusas, J. S. Duncan, A semisupervised joint network for simultaneous left ventricular motion tracking and segmentation in 4d echocardiography, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer (2020) 468–477.
22. Y. Chen, X. Zhang, C. M. Haggerty, J. V. Stough, Assessing the generalizability of temporally coherent echocardiography video segmentation, in: Medical Imaging 2021: Image Processing, Vol. 11596, International Society for Optics and Photonics (2021) 463–469.