GL-Fusion: Global-Local Fusion Network for Multi-view Echocardiogram Video Segmentation

Ziyang Zheng[1], Jiewen Yang[1], Xinpeng Ding[1], Xiaowei Xu[2(✉)], and Xiaomeng Li[1(✉)]

[1] The Hong Kong University of Science and Technology, Hong Kong SAR, China
eexmli@ust.hk

[2] Guangdong Cardiovascular Institute, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, Guangzhou, China
xiao.wei.xu@foxmail.com

Abstract. Cardiac structure segmentation from echocardiogram videos plays a crucial role in diagnosing heart disease. The combination of multi-view echocardiogram data is essential to enhance the accuracy and robustness of automated methods. However, due to the visual disparity of the data, deriving cross-view context information remains a challenging task, and unsophisticated fusion strategies can even lower performance. In this study, we propose a novel **G**obal-**L**ocal fusion (**GL-Fusion**) network to jointly utilize multi-view information globally and locally that improve the accuracy of echocardiogram analysis. Specifically, a **M**ulti-view **G**lobal-based **F**usion **M**odule (MGFM) is proposed to extract global context information and to explore the cyclic relationship of different heartbeat cycles in an echocardiogram video. Additionally, a **M**ulti-view **L**ocal-based **F**usion **M**odule (MLFM) is designed to extract correlations of cardiac structures from different views. Furthermore, we collect a multi-view echocardiogram video dataset (MvEVD) to evaluate our method. Our method achieves an 82.29% average dice score, which demonstrates a 7.83% improvement over the baseline method, and outperforms other existing state-of-the-art methods. To our knowledge, this is the first exploration of a multi-view method for echocardiogram video segmentation. Code available at: https://github.com/xmed-lab/GL-Fusion

Keywords: Multi-view fusion · Echocardiogram videos · Cardiac structure segmentation

1 Introduction

Accurate segmentation of the cardiac structure from echocardiogram videos is integral to several analysis tasks [11] and has a significant impact on clinical

Z. Zheng and J. Yang—Two authors contributed equally to this work.
Z. Zheng—Work completed during the internship at HKUST.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14223, pp. 78–88, 2023.
https://doi.org/10.1007/978-3-031-43901-8_8

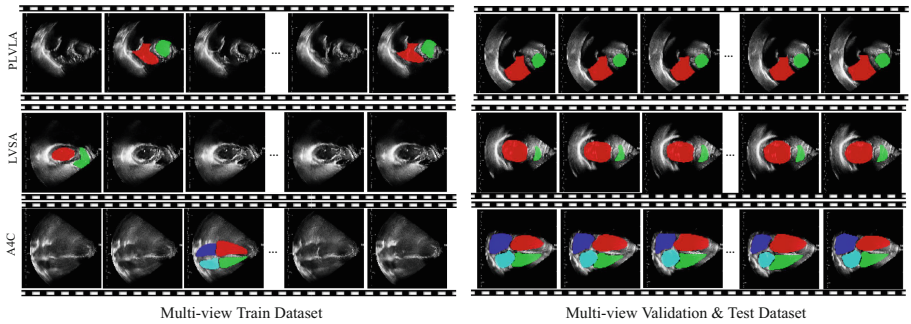

Fig. 1. Examples of multi-view echocardiogram dataset MvEVD, including PLVLA, LVSA, and A4C from top to bottom row. The colours red, green, blue, and cyan denote the LV, RV, LA, and RA cardiac structures. Our train set is sparsely annotated (5 frames per video), while the validation set and test set are fully annotated for each video frame. (Color figure online)

practice [26]. For example, segmentation of the left ventricle (LV) enables quantifiable functional analysis of the heart, facilitating the detection and diagnosis of heart diseases [3,20,21]. Compared with the single view segmentation, multi-view information is crucial to diagnose heart disease, *e.g.*, the diagnosis of congenital heart disease requires the analysis of four views: parasternal long-axis view (PSLAX), parasternal short-axis view (PSSAX), subxiphoid long-axis view (SXLAX), and suprasternal long-axis view (SSLAX) [22]. Consequently, to assist clinicians in diagnostic decision-making, there is a high demand for developing automated multi-view cardiac structure segmentation methods from echocardiogram videos in clinical practice. Existing echocardiogram segmentation approaches are primarily designed for single-view images or videos. For instance, Li et al. [26] proposed a dynamic neural network capable of segmenting the LV from a long-axis fetal echocardiogram. In comparison, Leclerc et al. [12] evaluated an encoder-decoder deep convolutional neural network that independently segments two and four-chamber images. However, these approaches have not addressed multi-view segmentation, where multi-view segmentation methods already exist in other medical domains, such as the CT-MRI [9,17,18], multi-view cardiac MRI [4,13,15,16], multi-view mammogram [2], and longitudinal multiple sclerosis [1]. Applying the proposed methods to multi-view echocardiogram segmentation presents several limitations: (1) Some methods are built for specific datasets and cannot adapt to our task. For instance, UMCT [25] designated supervised training in one view by generating pseudo segmentation labels from other views, but has limitations in our task due to the significant gaps between views. In contrast, InfoTrans [13] is designed for transmitting information between views instead of fusion them. While VCN [6] employs contrastive learning to predict volume but may not be suitable for our task since defining positive and negative pairs is challenging due to the significant gap between views and labels. (2) Methods such as JOIN [2], ROI-based fine-grained CNN [14],

MIMTP [1], MV U-Net [4], MV-CNN [23], and Type-I, II, III [9] concatenate the features or predicted probability maps of different views and then apply a fully-connected layer. However, these naive fusion strategies have shown limited performance and may even lead to worse results; see results in Table 1. (3) Existing multi-view segmentation methods such as TransFusion [15] and rDLA [16] mainly apply multi-view fusion with only global features. However, using global features for multi-view fusion may result in tangling the foreground/background pixels [10] or leads to high levels of background noise in echocardiograms.

To address this limitation, as shown in Fig. 1, we first collect a multi-view echocardiogram video dataset, including three views: parasternal left ventricle long axis (PLVLA view), left ventricular short axis (LVSA) view, and apical 4 chamber (A4C) view. Different views of echocardiograms contain annotations for different chambers, such as, the PLVLA view contains the left ventricle (LV) and right ventricle (RV), the LVSA view contains the LV and RV, and the A4C view contains the LV, left atrium (LA), right atrium (RA), and RV. Furthermore, we propose a novel global-local fusion (GL-Fusion) network for multi-view echocardiogram video segmentation, where GL-Fusion includes a multi-view local-global fusion module designed to aggregate information from different views and improve the representation of each view. The GL-Fusion comprises two components. First, a multi-view global fusion module (MGFM) interacts with the global semantics between different views and thus enhances the representation of each view. Second, since the global semantics may contain a significant amount of noisy information, a multi-view local fusion module (MLFM) is introduced to encourage the model to focus on foreground information.

In addition to capturing multi-view information, we propose a novel dense cycle loss designed to utilize unlabelled video data for improved representation learning. Our motivation is based on the idea that standard multi-view data is obtained from the same patient and under the same stable conditions, without abnormal behaviours such as suffocating or exercising, ensuring consistent cardiac cycles. Previous work [7] proposed an unsupervised method called cycle loss, which trains the model with unlabelled frames based on the heartbeat cycle's characteristics. Nevertheless, the proposed cycle loss only focuses on a pair in two different cycles but ignores possibly similar images that may appear simultaneously in a systolic or diastolic period, resulting in features from similar frames being considered distant. To address this issue, our dense cycle loss examines all possible pairings throughout the heartbeat cycle. In summary, our contributions are as follows:

- To the best of our knowledge, this is the first study to examine multi-view echocardiogram video segmentation.
- Our proposed GL-Fusion uses a multi-view local-global fusion module to combine information from different views and improve the representation of each view.
- We further design a dense cycle loss that utilizes unlabelled data to enforce feature similarity based on temporal cyclicality.

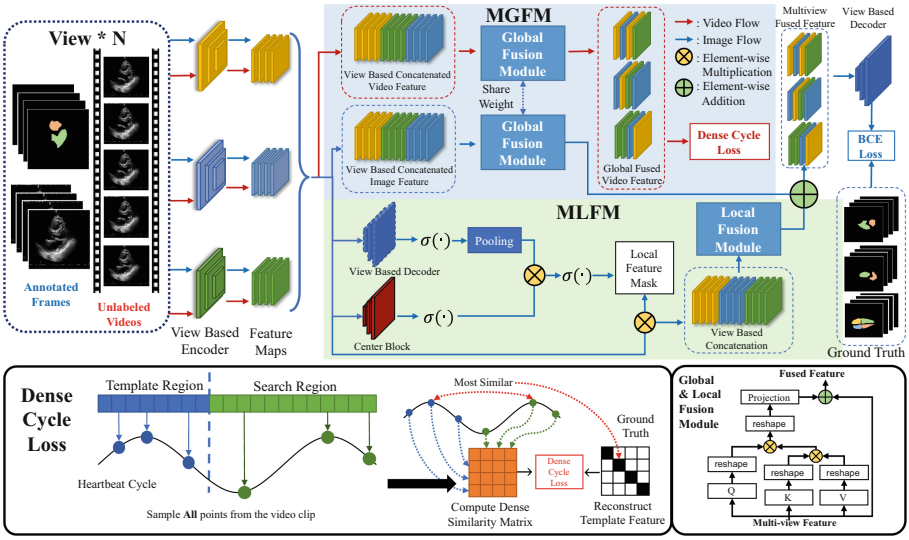

Fig. 2. The overview framework of GL-fusion. The Multi-view Global-based Fusion Module (MGFM) is proposed for global context information extraction and introduces dense cycle loss to devise the enforcement of the similarity of dense features between two heartbeat cycles from an echocardiogram video. The proposed Multi-view Local-based Fusion Module (MLFM) focuses on mining the correlation of local features of chambers in a different view.

– Extensive experiments demonstrate our method improved performance over existing methods, achieving an average dice score of 0.81. We plan to make our code publicly available upon paper acceptance.

2 Methodology

2.1 The Overall Framework

Figure 2 shows the overall pipeline of our proposed Multi-view Echocardiogram Global-Local Fusion Network (GL-Fusion), which consists of four main components: a view-based encoder, a multi-view global-local fusion module, a dense cycle loss module and a view-based decoder, where view-based indicate that parameters of the network of each independent view are non-shared. In our experiment, we use DeeplabV3 [5] as our view-based encoder and decoder.

Formally, we denoted the sample echocardiogram videos as $\mathbf{V} = \{\mathbf{X}^i\}_{i=1}^{V}$, where $\mathbf{X}^i \in \mathbb{R}^{C\times H\times W\times T}$ is the i-th view video and V is the number of views, and, C, H, W and T indicate the channels, height, width, and length of input images. Each video consists of T frames, *i.e.*, $\mathbf{X}^i = \{\mathbf{x}_t^i\}_{t=1}^{T}$, where T remain the same for different view and $\mathbf{x}_t^i \in \mathbb{R}^{C\times H\times W}$ indicate t-th frame of i-th view video.

Since only sparse frames are provided segmentation annotation for training in a video, thus we denote the annotation frame pair as $\{\mathbf{x}^i_{t_n}, \mathbf{y}^i_{t_n}\}_{n=1}^N$, where t_n is the index of the annotation and N is the number of labelled frames that $N << T$.

During the training, We feed the videos $\mathbf{V}$ into the view-based encoder to extract the corresponding feature maps $\{\mathbf{F}^i\}_{i=1}^V$ of each view, where $\mathbf{F}^i \in \mathbb{R}^{D\times h\times w\times T}$, and, D, h and w indicate the channel number, height and width of feature maps. Then the multi-view global-local fusion module aims to obtain the multi-view fused features $\{\overline{\mathbf{F}}^i\}_{i=1}^V$, which extract global and local semantics information from other views to enhance the representation of each view (See Sect. 2.2). Following is the view-based decoder that generates the predicted segmentation result $\mathbf{y}^i$ from fused features, and maps the results to corresponding segmentation annotation, *i.e.*, $\hat{\mathbf{y}}^i_{t_n}$ to the segmentation masks $\mathbf{y}^i_{t_n}$. For the annotated frames, we use the segmentation loss to supervise them, formulated as follows:

$$\mathcal{L}_{seg} = \sum_{i=1}^{V}\sum_{t_n=1}^{N} \mathcal{L}_{bce}(\hat{\mathbf{y}}^i_{t_n}, \mathbf{y}^i_{t_n}), \tag{1}$$

where $\mathcal{L}_{bce}$ is the Binary Cross Entropy. The sparse annotations are only a few frames in the whole video; thus can not obtain a robust model. To leverage a large number of unlabelled frames, we design the dense cycle loss $\mathcal{L}_{cyc}$ to enforce temporal feature similarity of videos based on cyclicality; See Sect. 2.3. The overall loss function of our model is as follows:

$$\mathcal{L} = \mathcal{L}_{seg} + \alpha\mathcal{L}_{cyc}, \tag{2}$$

where α is the hyper-parameter to control the weight between two losses. In the following, we will illustrate the multi-view global-local fusion module and the dense cycle loss in detail.

2.2 Multi-view Global-Local Fusion Module

In this section, we describe the multi-view global-local fusion module that aggregates the information from different views to enhance their feature representation. To this end, we first concatenate extracted feature $\{\mathbf{F}^i\}_{i=1}^V$ from different views in a view-wise manner to obtain $\mathbf{F} = \{\mathbf{f}_t\}_{t=1}^T$, where $\mathbf{f}_t$ is the t-th feature vector in $\mathbf{F}$, and $\mathbf{f}_t \in \mathbb{R}^{D\times V\times w\times h}$. Then, we describe the multi-view global and local fusion with $\mathbf{F}_{global}$ and $\mathbf{F}_{local}$, respectively.

Multi-view Global Fusion. In order to enhance the representation of each view, we propose the global-based fusion module (MGFM) to interact with the global semantics between different views. To this end, we introduce a view-wise non-local block, which extracts the context information across views. Similarly to the previous research [8,24] that applied attention to fuse the information, we here introduce the view-wise attention module to aggregate the cross-view information (see Fig. 2). Then fused feature $\overline{\mathbf{F}}_{global}$ will be sent to both compute the dense cycle loss and cooperate with the local fused feature for segmentation prediction.

Multi-view Local Fusion. Since each view represents different morphological information of the heart and may contain the same cardiac structure as others, for example, the view PLVLA and LVSA both contain left ventricle(LA) and right ventricle(RV). Hence, extracting the local feature that represents the cardiac structure can contribute to feature fusion more efficiently. In this module, the extracted feature $\mathbf{F}_{local}$ will first pass to both the view-based decoder and a center block, where the decoder and center block has the same components with different output. The decoder provides the pseudo label $\{\hat{\mathbf{y}}^i\}_{i=1}^V$ of different cardiac structures. A center block is introduced to acquire the weight $\{w^i\}_{i=1}^V$ of $\{\hat{\mathbf{y}}^i\}_{i=1}^V$ and compute the local feature masks $\{\mathcal{M}^i\}_{i=1}^V$ as Eq. 3,

$$\mathcal{M}^i = \sigma(pooling(\sigma(\hat{\mathbf{y}}^i)) \times \sigma(w^i)), \tag{3}$$

where weight w has the greatest volume in the central area of the segmented regions and attenuation with distance, σ denotes the sigmoid function and $\mathcal{M} \in \mathbb{R}^{1\times H\times W\times T}$. These masks highlight features with a stronger intensity that are closer to the object center, while discarding background information that is farther away from the center. This selection is based on the understanding that morphological information should remain consistent closer to the center. In the final, similar to the process of MLFM, the view-wise local feature will be conducted view-wise concatenation operation and multiplied with local feature mask $\{\mathcal{M}^i\}_{i=1}^V$. Then sent to the view-wise attention module to acquire the local fused feature $\overline{\mathbf{F}}_{local}$.

2.3 Dense Cycle Loss

In echocardiogram videos, since only sparse annotation is available for the supervised training, involving the unlabelled data for our training and enhancing the performance is a challenge. The previous research [7] proposes an unsupervised method named cycle loss, which jointly trains the model with the unlabelled frames according to the characteristic of the heartbeat cycle. However, the proposed cycle loss considers only one clip in an iteration, which has the possibility to match frames that are morphologically identical but not in the same state, such as the search region being end-diastole while the template region is end-systole.

Thus, we propose the dense cycle loss, which considers all the possible matching across all template and search regions in each view independently. For the multi-view fused feature $\overline{\mathbf{F}}_{global}$ of each video will be separated to template region P^i and search region Q^i with a ratio in 2:3 according to total frame length T. Then we densely sample all feature intervals $\{p_1^i, ..., p_n^i\}$ from P^i and $\{q_1^i, ..., q_m^i\}$ from Q^i, respectively, both sampling use the same chunk size s and in our experiment, n and m is $\frac{2}{5} \times \frac{T}{s}$ and $\frac{3}{5} \times \frac{T}{s}$. Then we compute the similarity between candidate interval p_k^i and target intervals q_j^i of Q^i.

$$\alpha_j^i = \sum \mathcal{W}(\{p\}_k^i, \{q\}_j^i) \times \{q\}_j^i, \tag{4}$$

Table 1. The comparison with other methods. all results are reported in Dice Score.

	Method	PLVLA	LVSA	A4C	Average Dice (%)
Single-view	DeeplabV3 [5]	70.93	75.14	77.33	74.46
	U-Net [19]	73.35	77.57	76.60	75.84
	CSS [7]	79.09	79.70	77.71	78.83
Fusion-based	Early-fusion	79.78	77.07	77.58	78.14
	Mid-fusion	77.89	76.75	72.44	75.69
	Late-fusion	71.62	75.31	74.68	73.87
	TransFusion [15]	78.79	80.23	59.31	72.78
	Ours	**83.84**	**81.76**	**81.28**	**82.29**

where $\mathcal{W}(\cdot)$ is the computation of the similarity matrix. The similarity will be used as the weight to reconstruct the feature interval $\tilde{p}_k^i$. Then we back to template region P^i and compute the similarity between $\tilde{p}_k^i$ and all feature intervals$\{p_1^i, ..., p_n^i\}$ in P^i. Then we consider the index of p_k^i as one-hot label g of the most similar interval of $\tilde{p}_k^i$ and compute view-wise cycle loss $\mathcal{L}_{cyc}$ with label g as shown in the following equation:

$$\mathcal{L}_{cyc} = \sum_{i=1}^{V} \sum_{j \in P^i} \mathbb{1}_{j=g} log(\alpha_j^i) \tag{5}$$

3 Experiment

Datasets. We collect a large multi-view echocardiogram video dataset named **MvEVD** from one medical institution, with a total of 254 sparsely annotated videos and 10 fully annotated videos with 800×600 resolution across three cardiac views (PLVLA, LVSA and A4C view). Each video includes 5 annotated frames. The average length of each video is larger than 100 frames that are able to cover more than one cardiac cycle.

Implementation Details. We use the model DeeplabV3 [5] as our view-based encoder and decoder, and select Adam optimizer for the model training with initial learning rate as $3e^{-4}$ and weight decay of $1e^{-5}$. When training, we use all sparsely annotated videos. All annotated frames are selected to supervise training while randomly selecting 40 consecutive frames from videos for semi-supervised training. The training batch size of annotated images and unlabeled videos is 8 and 1, respectively. In the final, we use CosineAnnealing as a scheduler and set the total training epoch to 100. The framework is built with Pytorch with 4 NVIDIA RTX3090 GPUs for training. For the data augmentation in the training stage, we resize each frame in 144×144 size and then randomly crop them to 112×112.

Table 2. Effectiveness of MGFM and MLFM. This table shows the performance of the Global and Local fusion modules

	MGFM	MLFM	Avg. Dice(%)
Base	✗	✗	74.46
Base+MGFM	✓	✗	80.20
Base+MLFM	✗	✓	78.41
Ours	✓	✓	**82.29**

Table 3. Effectiveness of Cyc. and Dense Cyc.. This table shows the effectiveness of vanilla cycle loss [7] (Noted by Cyc.) and our proposed dense cycle (Noted by Dense Cyc.).

	Cyc.	Dense Cyc.	Avg. Dice(%)
Fusion-only	✗	✗	80.36
Fusion+Cyc.	✓	✗	79.33
GL-Fusion	✓	✓	**82.29**

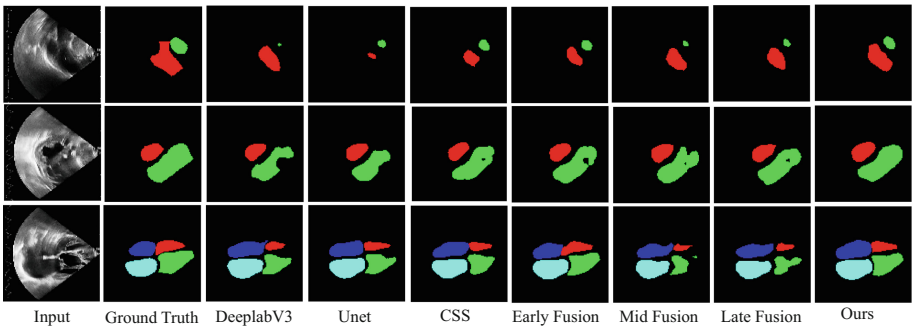

Fig. 3. Segmentation results from three views of echocardiogram videos, including PLVLA, LVSA, and A4C from top to bottom row. The red, green, blue, and cyan colours refer to LV, RV, LA, and RA cardiac structures, respectively. (Color figure online)

Validation and Testing. We use all fully annotated videos and split them into validation and testing with a ratio of 2:8. In this stage, we resize each frame in 144×144 size and conduct center cropping to them with the size of 112×112. Selecting the best model based on validation performance and report results in the testing set with Dice score.

3.1 Comparison with the State-of-the-Art Methods

To evaluate the performance of our method, we do the comparison with two types of methods: single-view methods and fusion-based methods in Table 1. To be specific, single-view methods independently train segmentation networks for each view without using any strategy across views or simply conducting semi-supervised approaches [7]. Fusion-based methods use feature-fusion modules to aggregate features and predict the segmentation masks. Our GL-Fusion method can reach 83.84%, 81.76% and 81.28% performance in Dice score across three different views, with 10.49%, 4.19% and 4.68% boosts when compared with the best single-view method [19], and 4.75%, 2.06%, 3.57% enhancement when compared to the best single-view with semi-supervised method CSS [7]. Also,

compared with the different global fusion methods, our global and local fusion methods conduct significant improvements compared with the early-fusion approach. The visualization in Fig. 3 compares the segmentation quality with our GL-fusion method and others across three different views.

3.2 Ablation Study

In this section, we analyze the contribution to the performance of the proposed modules Multi-view Global Fusion Module (MGFM) and Multi-view Global Fusion Module (MGFM) of our framework. All results are illustrated in Table **2**. a-b, the baseline without adapting any fusion strategy presents the lowest average dice, while using only MGFM or MLFM module can boost the result to 80.20% and 78.41%, respectively. The combination of these two modules can reach 82.29% dice score with a 2.09% increase in Dice score. In contrast, using the fusion method and cycle loss will lead to worse performance, while our proposed dense cycle loss can boost the result from 80.36% to 82.29%.

4 Conclusion

In this paper, we propose a novel fusion framework called GL-Fusion, which jointly uses global and local information to enhance the segmentation performance of echocardiogram videos. Additionally, to ensure fair evaluation of the multi-view segmentation results, we introduce a multi-view echocardiogram video dataset called **MvEVD**, which provides full annotation for validating and testing performance. Our results demonstrate that the proposed GL-Fusion framework significantly outperforms other methods. In the future, we aim to further improve our method and make it more efficient.

Acknowledgements. This work was partially supported by the Beijing Institute of Collaborative Innovation (BICI) under Grant HCIC-004, in collaboration with HKUST; the Foshan HKUST Projects under Grants FSUST21-HKUST10E and FSUST21-HKUST11E; and the Hong Kong Innovation and Technology Fund under Project ITS/030/21.

References

1. Birenbaum, A., Greenspan, H.: Longitudinal multiple sclerosis lesion segmentation using multi-view convolutional neural networks. In: Carneiro, G., et al. (eds.) LABELS/DLMIA -2016. LNCS, vol. 10008, pp. 58–67. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46976-8_7
2. Carneiro, G., Nascimento, J., Bradley, A.P.: Automated analysis of unregistered multi-view mammograms with deep learning. IEEE Trans. Med. Imaging **36**(11), 2355–2365 (2017). https://doi.org/10.1109/TMI.2017.2751523
3. Carneiro, G., Nascimento, J.C., Freitas, A.: The segmentation of the left ventricle of the heart from ultrasound data using deep learning architectures and derivative-based search methods. IEEE Trans. Image Process. **21**(3), 968–982 (2012). https://doi.org/10.1109/TIP.2011.2169273

4. Chen, C., Biffi, C., Tarroni, G., Petersen, S., Bai, W., Rueckert, D.: Learning shape priors for robust cardiac MR segmentation from multi-view images. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11765, pp. 523–531. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32245-8_58
5. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
6. Cheng, L.H., Sun, X., van der Geest, R.J.: Contrastive learning for echocardiographic view integration. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. MICCAI 2022. LNCS, vol. 13434. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16440-8_33
7. Dai, W., Li, X., Ding, X., Cheng, K.T.: Cyclical self-supervision for semi-supervised ejection fraction prediction from echocardiogram videos. IEEE Transactions on Medical Imaging (2022)
8. Ding, X., et al.: Support-set based cross-supervision for video grounding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11573–11582 (2021)
9. Guo, Z., Li, X., Huang, H., Guo, N., Li, Q.: Deep learning-based image segmentation on multimodal medical imaging. IEEE Trans. Radiat. Plasma Med. Sci. **3**(2), 162–169 (2019)
10. Hsu, C.-C., Tsai, Y.-H., Lin, Y.-Y., Yang, M.-H.: Every pixel matters: center-aware feature alignment for domain adaptive object detector. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12354, pp. 733–748. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58545-7_42
11. Hu, Y., et al.: Fully automatic pediatric echocardiography segmentation using deep convolutional networks based on biSeNet. In: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 6561–6564. IEEE (2019)
12. Leclerc, S., et al.: Deep learning for segmentation using an open large-scale dataset in 2D echocardiography. IEEE Trans. Med. Imaging **38**(9), 2198–2210 (2019)
13. Li, L., Ding, W., Huang, L., Zhuang, X.: Right ventricular segmentation from short- and long-axis MRIs via information transition. In: Puyol Antón, E., et al. (eds.) STACOM 2021. LNCS, vol. 13131, pp. 259–267. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-93722-5_28
14. Liang, S., Thung, K.H., Nie, D., Zhang, Y., Shen, D.: Multi-view spatial aggregation framework for joint localization and segmentation of organs at risk in head and neck CT images. IEEE Trans. Med. Imaging **39**(9), 2794–2805 (2020). https://doi.org/10.1109/TMI.2020.2975853
15. Liu, D., et al.: TransFusion: multi-view divergent fusion for medical image segmentation with transformers. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. MICCAI 2022. LNCS, vol. 13435. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16443-9_47
16. Liu, D., Yan, Z., Chang, Q., Axel, L., Metaxas, D.N.: Refined deep layer aggregation for multi-disease, multi-view & multi-center cardiac MR segmentation. In: Puyol Antón, E., et al. (eds.) STACOM 2021. LNCS, vol. 13131, pp. 315–322. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-93722-5_34
17. Patel, J.M., Parikh, M.C.: Medical image fusion based on multi-scaling (drt) and multi-resolution (dwt) technique. In: 2016 International Conference on Communication and Signal Processing (ICCSP), pp. 0654–0657. IEEE (2016)

18. Peiris, H., Chen, Z., Egan, G., Harandi, M.: Duo-SegNet: adversarial dual-views for semi-supervised medical image segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12902, pp. 428–438. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87196-3_40
19. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
20. Storve, S., Grue, J.F., Samstad, S., Dalen, H., Haugen, B.O., Torp, H.: Realtime automatic assessment of cardiac function in echocardiography. IEEE Trans. Ultrason. Ferroelectr. Freq. Control **63**(3), 358–368 (2016)
21. Tobon-Gomez, C., et al.: Benchmarking framework for myocardial tracking and deformation algorithms: an open access database. Med. Image Anal. **17**(6), 632–648 (2013)
22. Wang, J., et al.: Automated interpretation of congenital heart disease from multi-view echocardiograms. Med. Image Anal. **69**, 101942 (2021)
23. Wang, S., et al.: A multi-view deep convolutional neural networks for lung nodule segmentation. In: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 1752–1755 (2017). https://doi.org/10.1109/EMBC.2017.8037182
24. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision And Pattern Recognition, pp. 7794–7803 (2018)
25. Xia, Y., et al.: Uncertainty-aware multi-view co-training for semi-supervised medical image segmentation and domain adaptation. Med. Image Anal. **65**, 101766 (2020)
26. Yu, L., Guo, Y., Wang, Y., Yu, J., Chen, P.: Segmentation of fetal left ventricle in echocardiographic sequences based on dynamic convolutional neural networks. IEEE Trans. Biomed. Eng. **64**(8), 1886–1895 (2016)