Medical Image Analysis 78 (2022) 102397

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Semi-supervised segmentation of echocardiography videos via noise-resilient spatiotemporal semantic calibration and fusion

Huisi Wu[a,*], Jiasheng Liu[a], Fangyan Xiao[a], Zhenkun Wen[a], Lan Cheng[b], Jing Qin[c]

[a] Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China
[b] Department of Mathematics, The Hong Kong University of Science and Technology, Kowloon, Hong Kong
[c] Centre for Smart Health, School of Nursing, The Hong Kong Polytechnic University, Kowloon, Hong Kong

ARTICLE INFO

Article history:
Received 27 October 2021
Revised 14 January 2022
Accepted 18 February 2022
Available online 25 February 2022

Keywords:
Echocardiography
Video semantic segmentation
Deep learning
Temporal context extraction
Spatiotemporal semantic calibration
Spatiotemporal semantic fusion

ABSTRACT

We present a novel model for left ventricle endocardium segmentation from echocardiography video, which is of great significance in clinical practice and yet a challenging task due to (1) the severe speckle noise in echocardiography videos, (2) the irregular motion of pathological heart, and (3) the limited training data caused by high annotation cost. The proposed model has three compelling characteristics. First, we propose a novel adaptive spatiotemporal semantic calibration method to align the feature maps of consecutive frames, where the spatiotemporal correspondences are figured out based on feature maps instead of pixels, thereby mitigating the adverse effects of speckle noise in the calibration. Second, we further learn the importance of each feature map of neighbouring frames to the current frame from the temporal perspective so as to distinctively rather than uniformly harness the temporal information to tackle the irregular and anisotropic motions. Third, we integrate these techniques into the mean teacher semi-supervised architecture to leverage a large amount of unlabeled data to improve the segmentation accuracy. We extensively evaluate the proposed method on two public echocardiography video datasets (EchoNet-Dynamic and CAMUS), where the average dice coefficient on the left ventricular endocardium segmentation achieves 92.87% and 93.79%, respectively. Comparisons with state-of-the-art methods also demonstrate the effectiveness of the proposed method by achieving a better segmentation performance with a faster speed.

1. Introduction

In recent decades, significant progress has been made in imaging technologies for the diagnosis and treatment of cardiovascular disease (CVD), effectively reducing the mortality rate of CVD. Among these imaging modalities, echocardiography has been widely used in the diagnosis of various CVDs from heart failure to valvular heart diseases as it is convenient, cost-effective, and free of radiation (Chen et al., 2020a). Precise diagnosis of CVDs based on echocardiography depends on accurate segmentation of some key structures and measurement of some key metrics indicating cardiac functions. For example, a central metric of cardiac function is the left ventricular ejection fraction, which is used to diagnose cardiomyopathy, assess eligibility for certain chemotherapies, and determine indications for medical devices. The ejection fraction is formulated as the ratio of left ventricular end-systolic volume (ESV) and left ventricular end-diastolic volume (EDV), computed by (EDV - ESV) / EDV. Currently, experienced clinicians still manually annotate left ventricular endocardium to calculate the ejection fraction by subtracting the end-systolic volume from the end-diastolic volume. However, this procedure is tedious, time-consuming, and subjective. Besides, echocardiography segmentation is strongly influenced by the quality of data (Noble and Boukerroui, 2006). To the end, automatic echocardiography video segmentation is highly demanded clinically.

However, as shown in Fig. 1, it remains a challenging task as (1) echocardiography videos are full of speckle noise, (2) the movement of the pathological heart is often irregular, and (3) it is difficult to collect a large number of annotations for training. While some deep learning models (Baskaran et al., 2020; Leclerc et al., 2019a; Liu et al., 2021a; Moradi et al., 2019a; Smistad et al., 2020; Painchaud et al., 2020) have been proposed to segment cardiac substructures from echocardiographic videos, most of the these models conduct the segmentation task frame by frame and neglect the temporal consistency of cardiac motion between frames, which is crucial to improve both accuracy and efficiency. These networks, in general, are more suitable for 2D image segmentation tasks but

[*] Corresponding author.
E-mail address: hswu@szu.edu.cn (H. Wu).

https://doi.org/10.1016/j.media.2022.102397
1361-8415/© 2022 Elsevier B.V. All rights reserved.

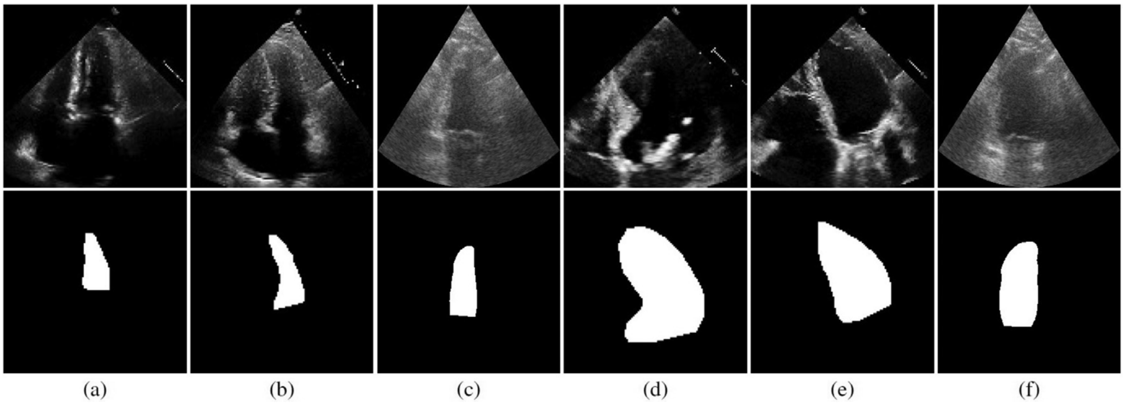

Fig. 1. Typical cases of echocardiography. For each video, there are only two frames (end-systole (a)–(c) and end-diastole (d)–(f)) are labeled.

not optimal choices for echocardiography video segmentation. In addition, in most benchmark datasets, for each echocardiography video, there are just a few frames that have been labeled due to the high annotation cost, which is also common in clinical practice. This makes the above-mentioned networks, which usually rely on a large number of labeled frames, more difficult to achieve satisfactory performance.

A straightforward idea is turning to modern natural video segmentation approaches (Hu et al., 2020; Jain et al., 2019), which have achieved remarkable performance in many natural video benchmark datasets, such as Cityscapes (Cordts et al., 2016), by harnessing the temporal consistency among frames. Unfortunately, these models cannot be directly applied to our task. On the one hand, the methods (Gadde et al., 2017; Ilg et al., 2017; Jain et al., 2019; Kroeger et al., 2016a; Nilsson and Sminchisescu, 2018; Sun et al., 2018; Zhu et al., 2017) based on optical flows are incapable of effectively capturing the motions in echocardiography videos due to (1) the irregular and anisotropic cardiac motions and (2) the existence of speckle noise, leading to performance degradation when dealing with echocardiography videos. In addition, the dense pixel-to-pixel mapping usually used in optical flows requires not only high computational cost but also is error-prone in ultrasound images as the pixel-wise mapping is susceptible to speckle noise. On the other hand, some models (Hu et al., 2020; Li et al., 2018; Shelhamer et al., 2016) attempt to reduce segmentation latency based on the principle of time consistency by reusing the deep semantic features of the previous frame and to support real-time or interactive segmentation. While in clinical practice, time performance is also significant in echocardiography video segmentation tasks, these models, with their limited representation capability, cannot achieve satisfactory performance because the extracted semantics are not that accurate due to the irregular cardiac motions and the limited labeled frames.

This paper proposes a novel model to comprehensively address these challenges by taking full advantage of spatiotemporal coherence among frames under speckle noise. We first design a temporal context-aware feature extraction module to make up for the important semantic information for segmentation that may be lost in subsequent operations by considering their temporal correlation. Then, based on these feature maps, we propose a novel adaptive spatiotemporal semantic calibration method to align these feature maps, in which the spatiotemporal correspondence is calculated based on the feature maps instead of pixels to generate a calibration convolution kernel, thereby mitigating the adverse effects of speckle noise in the calibration. Finally, we also learned a weight for each frame of the feature map to estimate its importance in ramping up the representative capability of the current frame according to their temporal coherence. We thus fuse and strengthen the current frame's feature map based on the learned weights by using a bi-directional spatiotemporal semantics fusion module.

In addition, inspired by some recently proposed semi-supervised segmentation models (Tarvainen and Valpola, 2017; Li et al., 2020; Wang et al., 2020; Liu et al., 2020), we integrate the proposed model into a mean teacher semi-supervised framework to leverage a large amount of unlabeled data to improve the segmentation accuracy with limited annotations. We conduct extensive experiments on two public echocardiography video datasets: EchoNet-Dynamic (Ouyang et al., 2020) and CAMUS (Leclerc et al., 2019b). Experimental results demonstrate the effectiveness of the proposed method, achieving better segmentation performance than state-of-the-art.

Our main contributions can be summarized as follows:

- We propose a novel semi-supervised model for left ventricle endocardium segmentation from echocardiography videos, which is capable of effectively taking both spatial and temporal information to improve segmentation performance under severe speckle noise and limited annotations.
- We propose a new adaptive spatiotemporal semantic calibration method that is able to adaptively align features of consecutive frames according to their importance based on feature maps instead of pixels, thereby greatly alleviating the adverse effects of speckle-noise on the calibration.
- We extensively evaluate the proposed method on two public datasets, achieving state-of-the-art performance for echocardiography video segmentation.

2. Related work

2.1. Echocardiography segmentation via deep learning

Cardiac ultrasound imaging (*a.k.a. echocardiography*) is an important and widely used clinical tool for evaluating various cardiovascular functions (Chen et al., 2020b). Due to the speckle noise and low quality of ultrasound images, echocardiography segmentation remains a very challenging problem. In recent years, deep learning based methods (Sheng et al., 2018; Liu et al., 2021b; Nazir et al., 2020; Ali et al., 2020; Wu et al., 2020; 2021a; 2021b) have replaced manually feature selection methods to improve accuracy and robustness of extracting features for image segmentation. According to the different forms of data applied in the segmentation network, these methods can be divided into echocardiographic image segmentation methods (Hu et al., 2019; Liu et al., 2021a; Ouyang et al., 2020; Leclerc et al., 2019a; Smistad et al., 2017; Veni et al., 2018) and echocardiographic video segmentation methods (Wei et al., 2020; Li et al., 2019; Chen et al., 2021b; Jafari et al., 2018; Pedrosa et al., 2017; Ta et al., 2020; Ahn et al., 2021).

The echocardiographic image segmentation strategy is a single-frame segmentation method with no correlation between frames. Hu et al. (2019) first proposed a deep learning method based on a bilateral segmentation network, where two segmentation paths are used to capture low-level spatial features and high-level context semantic features respectively to enhance the echocardiographic image segmentation performance. PLANet (Liu et al., 2021a) also proposed a pyramid local attention module, which further enhanced the feature extraction by capturing multi-scale information in a compact and sparse adjacent context, achieving even better echocardiographic image segmentation performance. However, the above-mentioned network usually relied on many labeled frames, ignoring the temporal consistency of heart motion between frames. Therefore, they cannot be directly applied in echocardiographic video segmentation.

To explore the temporal coherence between frames and improve the accuracy of echocardiogram segmentation by capturing additional temporal information, several representative echocardiogram video segmentation methods were also proposed by feeding a video or a sequence of images to the network for segmentation. Wei et al. (2020) first designed an appearance-level collaborative learning strategy, which simultaneously estimates the cardiac shape and motion based on two mutually benefited segmentation and tracking modules. By further designing another shape-level collaborative learning strategy, it can also enforce the temporal consistency according to the shape tracking results. Similarly, Li et al. (2019) also used a Conv-LSTM unit to analyze the spatial-temporal consistency between consecutive frames in the echocardiographic movie based on an optical flow motion estimation module to improve the segmentation accuracy. However, the high computational cost of the recursive unit hinders its real-time performance. In addition, the above methods still can only obtain a relatively poor accuracy in their motion estimations due to the challenging issues in ultrasound imaging, such as relatively low signal-to-noise ratios, varying speckle noise, and relatively low imaging contrast.

2.2. Video segmentation via spatiotemporal feature attention and fusion

Compared with single image data, video data contains temporal information. Therefore, in video segmentation, exploring the spatiotemporal information in the video is very important, where spatiotemporal feature attention and fusion technology is also commonly used to improve segmentation accuracy.

Based on the motion information of the segmentation targets, several methods tried to learn better feature representation ability by fusing the features between adjacent frames, which can be divided into three categories according to their approaches of modeling motion, including optical flow-based methods (Zhang et al., 2019; Jain et al., 2019; Nilsson and Sminchisescu, 2018; Gadde et al., 2017), the joint learning based methods (Lin et al., 2020; Chen et al., 2020c; Ding et al., 2020; Ta et al., 2020; Chen et al., 2021b), and the Conv-RNN based methods (Pfeuffer and Dietmayer, 2020; Wang et al., 2021a; Bai et al., 2018). Netwarp (Gadde et al., 2017) first employed a DIS-Flow (Kroeger et al., 2016b) to calculate the optical flow between adjacent frames, which can warp the shallow features of the previous frame to the corresponding position of the current frame to enhance the deep features of the current frame. Ding et al. (2020) proposed a new framework by jointing the video semantic segmentation and optical flow estimation, which improves the robustness of optical flow estimation based on the semantic segmentation of both occlusion and non-occlusion regions. Bai et al. (2018) further combined full convolutional networks and recurrent neural networks by incorporating spatial and temporal information into the segmentation task, which can solve the challenge of sparse video annotations by performing non-rigid label propagations with an exponentially weighted loss function. Although these methods can achieve higher accuracy than single-frame methods, their models are usually more complex. Their running times are still too long to achieve real-time performance in ultrasound video segmentation, especially for the optical flow or Conv-RNN with a relatively higher computational cost. Moreover, image-level modeling motion cannot handle the challenging noise, blur, and non-textured areas in ultrasound imaging.

On the other hand, several representative methods (Hu et al., 2020; Lu et al., 2019; Zhou et al., 2020; Ahn et al., 2021; Wang et al., 2021b; Lin et al., 2019; Wang et al., 2019; Hou et al., 2020) also explored the relationship between frames through the attention mechanisms. TDNET (Hu et al., 2020) first simulated the depth features by applying an attention propagation module to recombine the features extracted from the shallower sub-networks. Lu et al. (2019) also employed a global attention mechanism by jointing the feature spaces based on the common attention responses, which provides an effective stage for capturing global relevance and scene context. Ahn et al. (2021) further proposed a multi-frame attention network to improve the performance of left ventricular segmentation in three-dimensional echocardiography, where the multi-frame attention mechanism is established with highly correlated spatiotemporal features in the image sequence to improve video segmentation performance. Compared with optical flow based methods, the above methods can achieve a better segmentation effect through the attention mechanism at the feature level. However, the simple point-to-point attention mechanism can still not be directly applied in ultrasound images with severe noise. Instead, we can summarize the regional features of neighbouring frames to improve the feature fusion with a region-to-region manner, resulting in a significant accuracy improvement of echocardiographic video segmentation.

3. Methodology

The architecture of our proposed semi-supervised left ventricle endocardium segmentation network is illustrated in Fig. 2, which is implemented based on the mean teacher semi-supervised framework (Tarvainen and Valpola, 2017), mainly consists of temporal context-aware feature extraction and bi-directional spatiotemporal semantics calibration and fusion module. In addition, our semi-supervised network can be trained end-to-end and efficiently transmit labeled information from one labeled frame to the other unlabeled frames through supervised and unsupervised learning.

3.1. Temporal context-aware feature extraction

In most echocardiogram benchmark datasets, since each video has only two frames which are labeled on the end-systole and end-diastole, so we need fully exploit temporal context information to realize the important semantics transmission among different frames. Different from general video semantic segmentation approaches (Gadde et al., 2017; Jain et al., 2019), which either input the video to the network frame by frame without considering the temporal inter-frame coherence, or only repeatedly apply several frames to the same encoder and simply aggregate the extracted feature maps in the time domain, we propose a novel temporal context-aware feature extraction module to learn different feature representations of different relative time positions in the same frame.

By using k different independent encoders in the feature extraction among the neighbouring frames, we can maintain a better spatiotemporal consistency based on our adaptive semantic calibration and bi-directional semantics fusion modules. Specifically, given k consecutive frames in echocardiogram video as the input

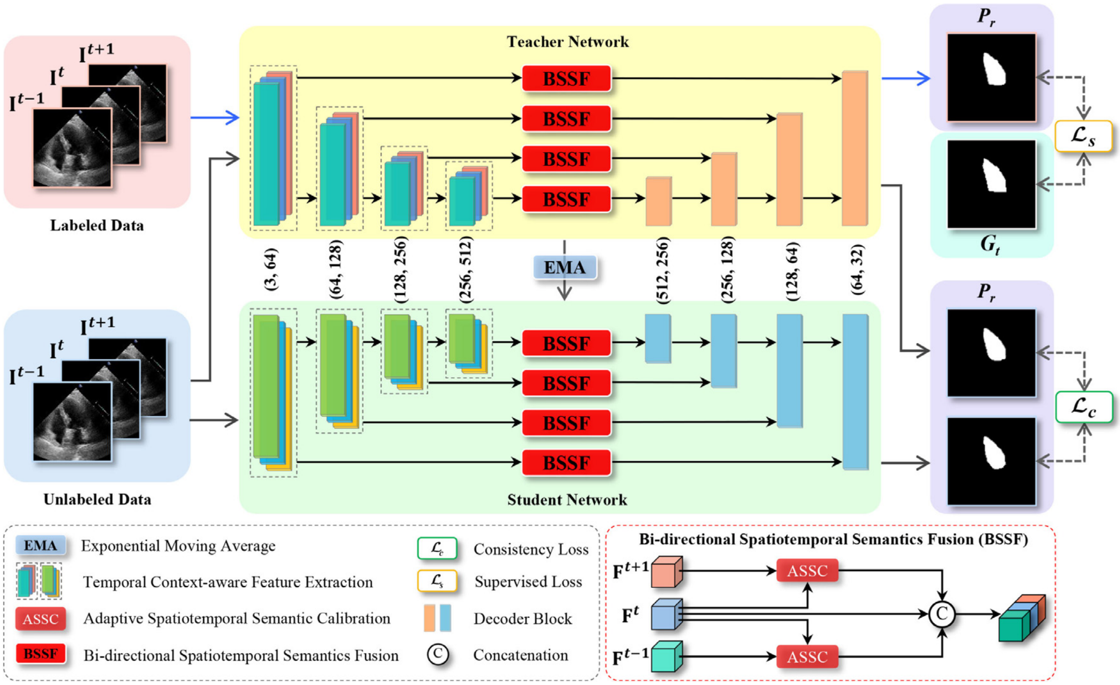

Fig. 2. Our proposed semi-supervised echocardiogram video segmentation network implemented based on the mean teacher semi-supervised framework mainly consists of a temporal context-aware feature extraction module and a bi-directional spatiotemporal semantics calibration and fusion module. EMA: exponential moving average.

for our network, as shown in Fig. 2, we employ k different encoders to learn feature representations independently, where the same frame in different relative positions will pass through different encoders. Each encoder has its own unique temporal context-aware weights through training. In this way, we can adaptively and explicitly define the semantic correlation between adjacent frames based on the temporal inter-frame coherence. Obviously, the running time may also significantly increase with the number for k. To simultaneously guarantee a real-time performance and a relatively better segmentation accuracy of the echocardiography videos, we finally choose $k = 3$ in our experiments.

Given 3 consecutive frames as the input $\{\mathbf{I}^{t-1}, \mathbf{I}^{t}, \mathbf{I}^{t+1}\} \in \mathbb{R}^{256\times256}$, we can obtain a set of feature maps on each layer $\{\mathbf{F}_i^{t-1}, \mathbf{F}_i^{t}, \mathbf{F}_i^{t+1}\} \in \mathbb{R}^{C_i\times H_i\times W_i}$ based on our proposed temporal context-aware encoders, where i denotes to the number of layers in each encoder, and H, W, C are the height, width, channel number of the feature map respectively. After temporal context-aware feature extraction, we can further perform the bi-directional spatiotemporal semantics calibration and fusion to improve the performance of semi-supervised echocardiogram video segmentation.

3.2. Adaptive spatiotemporal semantic calibration

Given the temporal context-aware feature maps, we can reinforce the left ventricular semantics features in the current frame based on a bi-directional spatiotemporal semantics fusion among the three neighbouring frames, which is crucial to the video segmentation performance. To obtain a precise semantics fusion between two neighbouring frames, there usually requires warping the feature maps from neighbouring frames to the target frame. Currently, most of existing feature warping technologies are implemented based on optical flows (Gadde et al., 2017; Jain et al., 2019; Nilsson and Sminchisescu, 2018; Zhu et al., 2017). Unfortunately, as mentioned before, due to the heavy ultrasound noise in the echocardiogram video and the relatively irregular and anisotropy cardiac motion, feature warping based on optical flow usually cannot obtain a high accuracy, which may inevitably decrease the following semantics fusion accuracy and eventually cut down the final video segmentation performance. More importantly, dense pixel-to-pixel mapping in the optical flow requires a much higher computational cost and is error-prone because the pixel-wise calculation is very sensitive to the ultrasound noise.

To overcome the above problem, we propose an adaptive spatiotemporal semantic calibration module to enhance the feature warping accuracy. As shown in Fig. 3, unlike the traditional optical flow that usually establishes the pixel-wise correspondence between the neighbouring frames, our adaptive spatiotemporal semantic calibration is implemented based on convolutions, where the spatiotemporal correspondences are calculated based on a set of convolutions upon several pixels. Obviously, based on a wider bandwidth signal to formulate the left ventricular motion between the neighbouring frames, our approach is potentially much more resistant to ultrasound noise. By training a series of sample-related convolution kernels between two adjacent feature maps, we can realize an adaptive spatiotemporal semantic calibration to enhance the following bi-directional spatiotemporal semantics fusion.

Specifically, to enhance the following bi-directional spatiotemporal semantics fusion, our adaptive spatiotemporal semantic calibration contains both coordinate warping calibration and channel-wise feature weighting calibration. Given two neighboring extracted feature map $\mathbf{F}_i^{t-1}$ and $\mathbf{F}_i^{t}$, we first apply two transformation matrixes $\mathbf{T}_q$ and $\mathbf{T}_k$ on $\mathbf{F}_i^{t-1}$ and $\mathbf{F}_i^{t}$ to obtain the feature map $\mathbf{Q} \in \mathbb{R}^{C_i\times H_i\times W_i}$ and feature map $\mathbf{K} \in \mathbb{R}^{C_i\times H_i\times W_i}$.

In our ASSC module, we assume that the spatial position offset between two consecutive frames usually does not exceed a few pixels due to the spatial-temporal consistency. Thus, we can formulate the transformation $\mathbf{T}_k$ and $\mathbf{T}_q$ as a 3×3 transformation matrix $\mathbf{T}$:

$$\mathbf{T} = \begin{bmatrix} r & s & t \\ u & v & w \\ 0 & 0 & 1 \end{bmatrix} \tag{1}$$

where r, s, t, u, v and w are 6 learnable parameters for the adaptive matrix. Obviously, the transformation $\mathbf{T}$ contains simple spatial

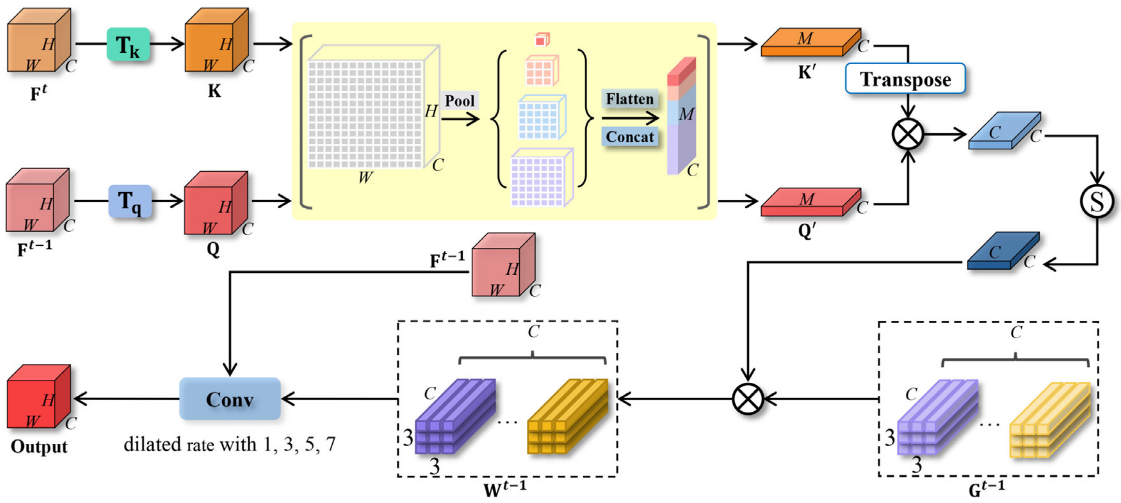

Fig. 3. Our proposed adaptive spatiotemporal semantic calibration module. The feature wrapping from adjacent frames to the target frame based on convolution instead of optical flow to resist the heavy ultrasound noise in the echocardiographic video and the relatively irregular and anisotropic heart motion.

transformations of translation, scaling, rotation, and their combination, which can also be implemented by a 3 × 3 convolution. To obtain an adaptive spatial transformation effect to minimize differences for the objects between two neighboring feature maps for subsequent operations, we can learn the above spatial transformation matrix **T** by jointly training with the total loss in our network.

On the other hand, we can further reduce the influence of speckle noise in our feature warping and semantics calibration by treating several pixels as a whole based on adaptive max-pooling with different kernel sizes on the **Q** and **K**, respectively. As shown in Fig. 3, we apply a series of max-poolings with different kernel sizes to summarize the spatial features in each channel separately and concatenate them in the flattened 1D dimension. Note that our adaptive spatial pooling and atrous spatial pyramid pooling (ASPP) (Chen et al., 2017) are similar in extracting the pyramid feature. However, there have obviously different purposes in designing our adaptive spatial pooling and the ASPP. By applying different dilated convolutions and concatenating the extracted features, ASPP is mainly used to expand the receptive field and improve the ability in extracting multi-scale features. Instead, our adaptive spatial pooling uses different pools to summarize and flatten the spatial features in each channel, which is mainly used to reduce the influence of speckle noise in feature distortion and semantic calibration.

By summarizing with the feature map based on the above max-pooling, we can minimize the influence of speckle noise and significantly reduce the computational cost in the following channel-wise feature weighting calibration. So far, we can obtain two summarized and rearranged feature maps $\mathbf{Q}' \in \mathbb{R}^{C_i \times M}$ and $\mathbf{K}' \in \mathbb{R}^{C_i \times M}$, which can be written as following,

$$\mathbf{Q}' = \mathbf{Summarize}(\mathbf{T}_q(\mathbf{F}_i^{t-1})), \mathbf{K}' = \mathbf{Summarize}(\mathbf{T}_k(\mathbf{F}_i^{t})) \tag{2}$$

where **Summarize** represents the spatial feature summarization based on multiple adaptive max-pooling with different kernel sizes. In our experiments, we employ four adaptive max pooling to summarize the spatial features to generate feature map with sizes of 1 × 1, 3 × 3, 5 × 5, and 7 × 7, where spatial resolution of $\mathbf{F}_i^{t-1}$ and $\mathbf{F}_i^{t}$ are summarized from $H_i \times W_i$ to $M = (1^2 + 3^2 + 5^2 + 7^2) = 84$.

To realize the channel-wise feature weighting calibration, we first obtain the channel-wise feature correlation between the two neighbouring frames via multiplying the summarized feature map $\mathbf{Q}'$ with the corresponding neighbouring feature map $(\mathbf{K}')^T$, where T represents a matrix's transpose. Based on the channel-wise feature correlation between the two neighbouring frames, we can form attention weights by applying sigmoid to the correlation matrix, and use the attention weights to aggregate motion warping matrix $\mathbf{G}_i^{t-1}$. We can obtain the adaptive spatiotemporal semantic calibration kernels $\mathbf{W}_i^{t-1} \in \mathbb{R}^{C_i \times C_i \times 3 \times 3}$ related to the upper and lower frames, written as,

$$\mathbf{W}_i^{t-1} = \mathbf{Sigmoid}(\mathbf{Q}' \times (\mathbf{K}')^T) \times \mathbf{G}_i^{t-1} \tag{3}$$

where **G** is a globally shared trainable motion warping matrix between two neighbouring frames. In our model, there are $2 * i$ **G**, namely $\mathbf{G}_i^{t-1}$ and $\mathbf{G}_i^{t+1}$ (representing the relationship matrix between the previous frame and the current frame and the relationship matrix between the next frame and the current frame in the ith layer). Specifically, we define it in advance (similar to hyper-parameters) before the training stage, and then perform temporal consistency learning as the training is affected by backpropagation. For the initialization of **G**, we define it as all 1, which have the same kernel size and input/output dimensions as the calibration kernel.

Finally, after the adaptive spatiotemporal semantic calibration kernel is obtained, we use the same semantic calibration block to calibrate the offsets among different frames based on the spatiotemporal consistency. As shown in Fig. 3, our semantic calibration block contains calibration kernels with different dilated rates to adaptively calibrate different motion offsets between neighbouring frames. Thus, we can warp the feature map of the previous frame to the current frame to enhance the semantic features of the current frame.

Note that, by applying our adaptive spatiotemporal semantic calibration kernels $\mathbf{W}_i^{t-1}$ on $\mathbf{F}_i^{t-1}$, simultaneously realizes both coordinate warping calibration and channel-wise feature weighting calibration, where several regional pixels are involved in reducing the influence of speckle noise. More importantly, the kernel parameters are fully determined by training based on the input neighbouring feature frames. Therefore, we can achieve a better bi-directional spatiotemporal semantics fusion after adaptive spatiotemporal semantic calibration.

3.3. Bi-directional spatiotemporal semantics fusion

In the above adaptive spatiotemporal semantic calibration, we use multiple pooling operations to treat multiple pixels as a whole to reduce the influence of speckle noise, which is proved to be effective. However, it will inevitably lose a small part of important information used for segmentation, such as edge information.

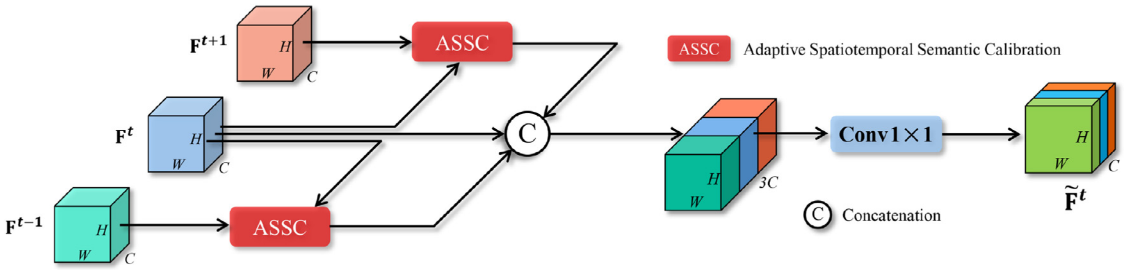

Fig. 4. Bi-directional spatiotemporal semantics fusion. Based on the adaptive Spatiotemporal semantic calibration module, the semantic features of adjacent frames are warping, and then seamlessly fused with the semantics of the current frame.

Therefore, we can fully utilize the spatiotemporal coherence based on a bi-directional spatiotemporal semantic fusion to compensate for the semantic loss and enhance the current frame features to improve our video segmentation performance. As shown in Fig. 4, given three consecutive neighboring feature maps $\{\mathbf{F}_i^{t-1}, \mathbf{F}_i^t, \mathbf{F}_i^{t+1}\}$, we first obtain two set of bi-directional adaptive spatiotemporal semantic calibration kernels $\mathbf{W}_i^{t-1}$ and $\mathbf{W}_i^{t+1}$ based on above semantics calibration method. As a result, our bi-directional spatiotemporal semantics fusion can be formulated as follows,

$$\tilde{\mathbf{F}}_i^t = Conv_{1\times1}(\{\mathbf{F}_i^{t-1} * \mathbf{W}_i^{t-1}, \mathbf{F}_i^t, \mathbf{F}_i^{t+1} * \mathbf{W}_i^{t+1}\}) \tag{4}$$

where W_i^{t-1} and W_i^{t+1} are trained through forward semantic calibration between $\{\mathbf{F}_i^{t-1}, \mathbf{F}_i^t\}$ and backward semantic calibration between $\{\mathbf{F}_i^t, \mathbf{F}_i^{t+1}\}$. After bi-directional spatiotemporal semantics fusion is conducted in each layer, we simply employ the well-validated U-Net decoder structure to propagate the fused semantic information to higher resolution layers obtain dense predictions, as shown in Fig. 2.

3.4. Loss functions

As shown in Fig. 2, we apply our proposed network to echocardiographic video segmentation under a mean teacher framework, where the structure of the teacher network and the student network are the same. However, their parameters are not shared, where the parameters between them are transferred by an exponential moving average (EMA) method. Note that our semi-supervised network training is performed end-to-end, which includes a supervised training stage that uses labeled frames to only guide the teacher network, and an unsupervised consistency training stage that uses unlabeled frames to guide the teacher and student networks.

3.4.1. Supervised training

For the labeled frames, we combine each labeled frame with its adjacent frames to input our teacher network, where we employ the labeled mask as the ground truth of segmentation. Cross-entropy loss is the most commonly used loss function task for image semantic segmentation, which compares the prediction result of each pixel category with the label vector. As our segmentation result has only contained two classes, given the prediction P_r and the ground truth G_t, we can employ a binary cross-entropy loss as a component in our loss function,

$$\mathcal{L}_{bce} = -\sum_i (g_i \ln(p_i) + (1-g_i)\ln(1-p_i)) \tag{5}$$

In addition, considering that the distributions of LV_{endo} in our segmentation result can be very irregular, to minimize the bias, we also follow Eelbode et al. (2020) to use a Dice loss as another component in our loss function,

$$\mathcal{L}_{dice} = 1 - \frac{2\sum_{i=1}^N p_i g_i}{\sum_{i=1}^N p_i^2 + \sum_{i=1}^N g_i^2} \tag{6}$$

where N is the total pixel numbers. $p_i \in [0,1]$ and $g_i \in \{0,1\}$ denote the pixel values at the ith position in P_r and G_t respectively.

Finally, we can formulate the total supervised loss $\mathcal{L}_s$ with three components, including dice loss $\mathcal{L}_{dice}$, binary cross-entropy loss $\mathcal{L}_{bce}$, and an L2 regularization term. Thus, our supervised loss $\mathcal{L}_s$ can be formulated as following,

$$\mathcal{L}_s = \alpha \cdot \mathcal{L}_{dice}(P_r, G_t) + \beta \cdot \mathcal{L}_{bce}(P_r, G_t) + \frac{\lambda}{2} \cdot \|\omega\|_2^2 \tag{7}$$

In supervised training stage, we set the weights $\alpha = 0.5$, $\beta = 0.5$ and $\lambda = 10^{-6}$ in our experiments.

3.4.2. Unsupervised training

For the unlabeled frames, we also combine each frame with the same number of adjacent frames to input our semi-supervised framework. By feeding them into the student network and teacher network, respectively, we can obtain two prediction results, including student network prediction P_r^s and teacher network prediction P_r^t. Here, we need to formulate the consistency loss $\mathcal{L}_c$ to standardize and constrain our segmentation modules. Similar to most semi-supervised networks, we can employ the widely used mean square error loss $\mathcal{L}_{mse}$ and the famous Kullback Leibler divergence loss $\mathcal{L}_{kl}$ in our consistency loss $\mathcal{L}_c$, which can be written as follows,

$$\mathcal{L}_{mse} = \frac{1}{n}\sum_{i=1}^N \left(p_i^s - p_i^t\right)^2 \tag{8}$$

$$\mathcal{L}_{kl} = D_{KL}(P_r^s \| P_r^t) = \sum_{i=1}^N p_i^s \cdot \left(\log p_i^s - \log p_i^t\right) \tag{9}$$

$$\mathcal{L}_c = \alpha \cdot \mathcal{L}_{mse}\left(P_r^s, P_r^t\right) + \beta \cdot \mathcal{L}_{kl}\left(P_r^s, P_r^t\right) \tag{10}$$

where N is the total pixel numbers. $p_i^t \in [0,1]$ and $p_i^s \in [0,1]$ denotes the ith position in P_r^t and P_r^s respectively. In unsupervised training stage, we set the weights $\alpha = 0.5$, and $\beta = 0.5$ in our experiments.

3.4.3. Parameter learning

The total loss $\mathcal{L}_{total}$ of our network can be formulated as,

$$\mathcal{L}_{total} = \sum_{i=1}^N \mathcal{L}_s(x_i) + \lambda \sum_{j=1}^U \mathcal{L}_c\left(y_j\right) \tag{11}$$

where N and U are the numbers of labeled frames and unlabeled frames. $\mathcal{L}_s(x_i)$ denotes the supervised loss for the ith labeled frame, while $\mathcal{L}_c(y_i)$ is the consistency loss for the jth unlabeled frame. The weight λ is used to balance the supervised loss on labeled data and the consistency loss on unlabeled data. In our experiments, we use a time-correlated function to update λ: $\lambda(t) = (\frac{t}{t_{\max}})^2$, where $\mathbf{t}$ denotes the current training iteration and $\mathbf{t}_{\max}$ is the maximum training iteration.

After formulating the total loss $\mathcal{L}_{total}$ for each batch, we still need to perform backward propagation to train the student network. During backward propagation training, we update

Table 1
Detail settings of the datasets.

Datasets	EchoNet-Dynamic	CAMUS
Patients	10,030	500
Available	10,030	450
Modality	2D Ultrasound	2D Ultrasound
Views	AC4	A2C, A4C
Sequences	10,030	450 + 450
Frames(avg)	120	20
FPS	50	-
Resolution	112 × 112	778 × 594
Labeled	20,060	1800
Position	ES, ED	ES, ED
Structure	LV_{endo}	LV_{endo}, LV_{epi}, LA

the teacher network parameters in each training step through the exponential moving average (EMA) strategy in Tarvainen and Valpola (2017). The parameters of teacher network in the tth training iteration are written as,

$$\theta'_t = \eta\theta'_{t-1} + (1-\eta)\theta_t \tag{12}$$

where θ_t is the student network parameter at the t training iteration. Similar to Tarvainen and Valpola (2017), we also set the value of EMA decay rate η as 0.99 during backward propagation training.

3.5. Implementation details

We implemented our proposed network based on pytorch1.6 (Paszke et al., 2017) and trained our network on a single GTX 2080Ti GPU, where the network parameters are randomly initialized and trained from scratch. In our experiments, we set the initial learning rate as 10^{-4}. We employ Adam to optimize the network during 100 epochs. During training process, we use the poly strategy to update learning-rate $l : l(t) = l(t-1)\cdot\left(1-\frac{iter}{total}\right)^{power}$, where $power = 0.9$. From the beginning frame in each echocardiogram video, we always take three consecutive frames as the input for model training. Based on our bi-directional spatiotemporal semantics calibration and fusion, we can fully utilize all unlabeled frames in our network training to enhance the echocardiogram video segmentation performance. In the test stage, we also employ the above sliding window mode to take every three consecutive frames as the input and feed them only into the student network. We can obtain the prediction as the final output of our proposed method. Note that we do not further use any techniques to post-process or refine the prediction results, which also saves a lot of computation cost to meet the real-time speed requirements for the echocardiogram video segmentation tasks.

4. Experiments

4.1. Datasets

To evaluate the effectiveness of our proposed semi-supervised echocardiogram video segmentation network, we performed the experiments on the two public echocardiographic datasets, including the latest published EchoNet-Dynamic (Ouyang et al., 2020) dataset and the widely used CAMUS (Leclerc et al., 2019b) dataset. The detailed settings of the datasets are as shown in Table 1. EchoNet-Dynamic is a large dataset containing 10,030 video sequences, including cardiac patients and healthy volunteers. Each echocardiogram video has about 200 frames with a resolution of 112 × 112 pixels, where only two frames (end-systole and end-diastole) are manually labeled by the experienced cardiologist. The CAMUS dataset includes 2D apical four-chamber and two-chamber sequences collected from 450 patients. Each sequence has about 20 frames with a resolution of 778 × 594 pixels, which also only has two manually annotated frames (end-systole and end-diastole). To evaluate the performance of our proposed method and its generalization ability, we employed a 5-fold cross-validation method in our experiments. For EchoNet-Dynamic and CAMUS datasets, we use 5-fold cross-validation based on an 80/20 split during training/validation. Specifically, we take the 8024 and 360 ultrasonic image sequences as training data and the next 2006 and 90 sequences as test data, respectively. We repeat the above process five times until the entire dataset is covered, where we can ensure that the same patient will not be in the same fold. In this way, we can reduce the deviation as much as possible and ensure the reliability of the experimental results. To obtain better efficiency, we resized both labeled and unlabeled images in the two echocardiogram video datasets to a uniform resolution of 256 × 256. To avoid overfitting and improve the generalization of our model among different echocardiogram video datasets, we also performed four kinds of data augmentations to enrich the diversity of samples, including horizontal flipping, vertical flipping, diagonal flipping, and random rotation with a degree between [-15°, 15°].

4.2. Evaluation metrics

In our experiments, we used geometric and clinical metrics to evaluate the performance of our proposed method for the segmentation in echocardiographic videos. The geometric indicators mainly include Dice coefficient (DC) and Hausdorff distance (HD), which be written as:

$$\mathrm{DC} = \frac{2\times|A\cap B|}{|A|+|B|} = \frac{2\times TP}{2\times TP+FP+FN} \tag{13}$$

$$\begin{aligned}\mathrm{HD} &= \max\{d_{AB}, d_{BA}\}\\ &= \max\left\{\max_{a\in A}\min_{b\in B} d_{(a,b)}, \max_{b\in B}\min_{a\in A} d_{(a,b)}\right\}\end{aligned} \tag{14}$$

where A is the predicted set of pixels; B is the ground truth of the segmentation, TP means the true positive, FP means the false positive, and FN means the false negative. Note that we only evaluate the two labeled frames for each sequence. Additionally, the area under the curve (AUC) of the receiver operating characteristic curve (ROC) is also employed to evaluate our model based on the recall and precision to measuring the segmentation performance.

Based on the ground truth and the prediction, we can also calculate the correlation coefficient (corr), average deviation (bias), and standard deviation (std) of left ventricular ejection fraction ($\mathrm{LV_{EF}}$). The calculation formula of left ventricular ejection fraction ($\mathrm{LV_{EF}}$) can be written as follows:

$$\mathrm{LV_{EF}} = \frac{LV_{EDV} - LV_{ESV}}{LV_{EDV}} \times 100\% \tag{15}$$

where LV_{EDV} and LV_{ESV} represent the volume of end-diastole (ED) and the volume of end-systole (ES), respectively.

4.3. Ablation studies

To demonstrate the improvement of our proposed method, we conducted the ablation studies to assess the importance of equipping both temporal context-aware encoder (TCE) and bi-directional spatiotemporal semantics fusion (BSSF) modules in our semi-supervised echocardiogram video segmentation. In our experiments, we implemented a baseline method by embedding an UNet on the mean teacher semi-supervised framework. By adding temporal context-aware encoder and bi-directional spatiotemporal semantics fusion module to the Baseline, respectively, we can obtain two approaches named Baseline+TCE and Baseline+BSSF. And Baseline+TCE+BSSF, the complete proposed method, is obtained by simultaneously adding the temporal context-aware encoder and bi-directional spatiotemporal semantics fusion module to the baseline network.

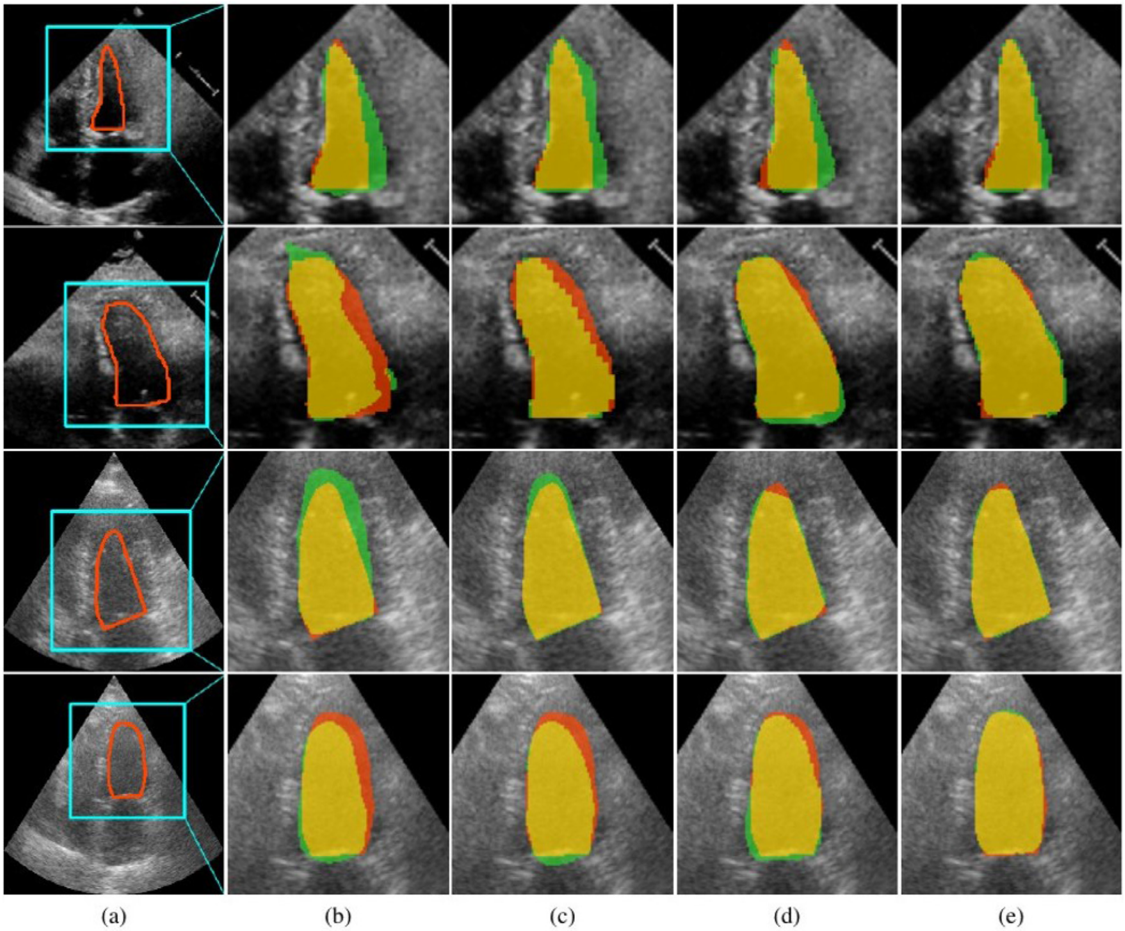

Fig. 5. Visual comparison of our ablation studies. (a) Input image. (b) Baseline. (c) Baseline+TCE. (d) Baseline+BSSF. (e) Baseline+TCE+BSSF. The red, green, and yellow regions represent the ground truth, prediction, and overlapping regions, respectively. TCE: temporal context-aware encoder. BSSF: bi-directional spatiotemporal semantics fusion.

Typical echocardiogram segmentation results in both the EchoNet-Dynamic dataset and the CAMUS dataset are as visualized in Fig. 5. By simply embedding an UNet on the mean teacher semi-supervised framework and roughly fusing three continuous frames during the echocardiogram video segmentation, the Baseline method cannot always accurately segment the left ventricular, especially for the ambiguous boundary with heavy speckle noise (Fig. 5(b)). Relying on a better temporal context-aware encoder, Baseline+TCE obtained better left ventricular segmentation performance than the Baseline, indicating the effectiveness of learning temporal context-aware feature representations for different frames (Fig. 5(c)). By adding the bi-directional spatiotemporal semantics fusion module, we can easily observe that the Baseline+BSSF method further outperformed both Baseline and Baseline+TCE method in echocardiogram video segmentation (Fig. 5(d)), where the segmentation for challenging cases with relatively blurred boundary regions is still not very stable, maybe due to the loss of part of the important semantic information used for segmentation, such as boundary information. By combining both TCE and BSSF modules in the baseline network, our method generally achieved much more accurate echocardiogram video segmentation performance. It shows that the TCE module has a supplementary effect on the part of the semantics lost by the BSSF module for extreme cases. As shown in Fig. 5(e), even for the ambiguous boundaries with heavy ultrasonic noise, our method still can obtain relatively accurate left ventricular segmentation results, which is almost overlaps with the ground truth regions. In addition, we also performed a statistical comparison by collecting the mean DC, HD, and AUC values for different methods on the EchoNet-Dynamic dataset and the CAMUS dataset, respectively. As shown in Table 2, we can clearly observe the performance of just a U-Net without the mean-teacher strategy training is slightly worse than that of Baseline, demonstrating that a large amount of unlabeled data is beneficial and should be fully utilized for echocardiographic segmentation. We can clearly observe the effectiveness of TCE and BSSF modules by comparing the Baseline+TCE and Baseline+BSSF with the Baseline method, respectively. On the other hand, it is also evident from Table 2 that our method generally outperformed other methods with relatively higher performance for all three evaluation metrics, implying that integrating both TCE and BSSF modules is important and efficacious for ventricular endocardial segmentation.

4.4. Comparison with state-of-the-art methods

To further demonstrate the advantages of our method in the challenging echocardiogram video segmentation, we also compared our proposed network with eight state-of-the-art methods, including three most representative networks for natural video segmentation (DFF (Zhu et al., 2017), Accel (Jain et al., 2019) and TDNet (Hu et al., 2020) and five latest published medical video segmentation networks (Joint-motion Qin et al., 2018, C-LSTM Bai et al., 2018, MFP-Net Moradi et al., 2019b, Joint-net (Ta et al., 2020), TransUNet (Chen et al., 2021a), and PLANet (Liu et al., 2021a). Among them, Joint-motion is a 2.5D

Table 2
Statistical comparison of our ablation studies. The U-Net is trained without using the mean-teacher strategy. TCE: temporal context-aware encoder. BSSF: bi-directional spatiotemporal semantics fusion.

Method	EchoNet-Dynamic			CAMUS		
	DC (%)	HD (mm)	AUC (%)	DC (%)	HD (mm)	AUC (%)
U-Net	89.34 ± 0.41	4.36 ± 1.13	95.21 ± 0.26	90.73 ± 0.45	3.94 ± 1.07	95.54 ± 0.14
Baseline	90.67 ± 0.22	3.92 ± 0.92	96.13 ± 0.19	91.81 ± 0.33	3.87 ± 1.21	96.68 ± 0.09
Baseline+TCE	91.69 ± 0.37	3.51 ± 1.18	97.38 ± 0.28	92.83 ± 0.41	3.49 ± 0.71	97.87 ± 0.27
Baseline+BSSF	91.33 ± 0.21	3.27 ± 0.72	97.15 ± 0.36	92.65 ± 0.27	3.15 ± 0.87	97.59 ± 0.21
Baseline+TCE+BSSF	**92.87 ±0.16**	**2.93 ±0.72**	**98.20 ±0.21**	**93.79 ±0.26**	**2.86 ±0.66**	**98.52 ±0.23**

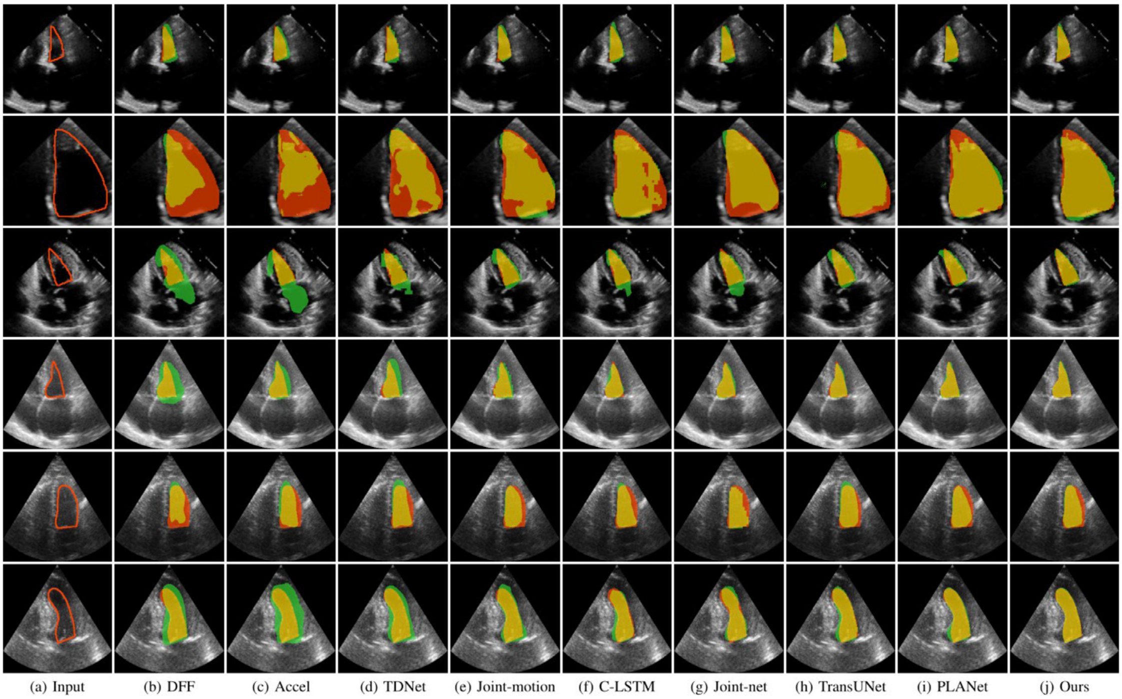

Fig. 6. Visual comparison with different state-of-the-art methods on the EchoNet-Dynamic and CAMUS datasets. Red, green, and yellow regions represent the ground truth, perdition and the overlapping region between them, respectively.

approach, C-LSTM is based on the long short-term memory network, and the other three are naive 2D approaches.). To guarantee a fair comparison, we have implemented all eight competitors and our proposed method with the 5-fold cross-validation under the same computational environments with the same data augmentations and the same learning rate adjustment settings.

From the typical echocardiogram frames shown in Fig. 6, we can see several challenging cases with various scales and irregular shapes of different left ventricular. Although optical flow-based networks have achieved excellent performance in natural video segmentation, DFF, Accel and TDNet still cannot achieve satisfying results for echocardiogram video segmentation, as shown in Fig. 6(b), (c) and (d). By learning a joint multi-scale feature encoder and estimating the LV motion under a weakly-supervised framework, the Joint-motion method is suitable for cardiac MR sequence segmentation. However, it still cannot accurately segment left ventricular from echocardiogram video with much more spackle noise and relatively sparse annotation (Fig. 6(e)). Similarly, the C-LSTM method cannot transform the performance from the cardiac MR sequence to echocardiogram video segmentation by incorporating both spatial and temporal information into the recurrent neural network. Based on carefully smoothing on displacement map between a source and a target frame, Joint-net outperformed both Joint-motion and C-LSTM methods, as shown in Fig. 6(f) and (g). More recently, by applying transformers for sequence-to-sequence prediction or a deep pyramid local attention neural network, TransUNet and PLANet also obtained some improvements in echocardiogram video segmentation. However, they still cannot handle well the challenging cases with more complex scale, shape variations, and blur echocardiogram boundaries), as shown in Fig. 6(h) and (i). As shown in Fig. 6(j), our method generally outperformed the other competitors by achieving better left ventricular segmentation results, which are much closer to the ground truth. Even for the challenging cases with many ambiguous boundaries, our results are still very close to the ground truth regions.

Besides visual comparisons, we also performed a statistical comparison for different methods on the EchoNet-Dynamic dataset and the CAMUS dataset, respectively. As shown in Tables 3 and 4, we can observe that natural video segmentation methods (DFF, Accel, and TDNet) generally fall behind the other latest medical video segmentation networks (Joint-motion, C-LSTM, Joint-net, TransUNet, and PLANet) about 1%-2% in terms of DC, HD, and AUC metrics. Obviously, it is evident from Tables 3 and 4 that our method generally outperformed other competitors with the highest performances in DC, HD, and AUC metrics for both EchoNet-Dynamic and CAMUS datasets. By integrating both TCE and BSSF modules in our method, we can not only minimize the influence of speckle noise to improve the accuracy performance but also signif-

Table 3
Statistical comparison with different state-of-the-art methods on the Echo dataset. Numbers format: mean value ± (standard deviation).

Method	Year	DC (%)	HD (mm)	AUC (%)
DFF	2017	85.63 ± 0.36	4.52 ± 1.19	93.64 ± 0.29
Accel	2019	86.78 ± 0.41	4.42 ± 1.14	94.20 ± 0.37
TDNet	2021	89.37 ± 0.35	3.87 ± 0.98	95.08 ± 0.24
Joint-motion	2018	90.97 ± 0.46	3.63 ± 0.77	96.75 ± 0.43
C-LSTM	2018	90.27 ± 0.25	3.75 ± 0.90	96.66 ± 0.37
Joint-net	2020	90.91 ± 0.36	3.85 ± 0.92	96.14 ± 0.39
Echo-Net	2020	91.50 ± 0.29	3.55 ± 0.73	96.64 ± 0.38
TransUNet	2021	91.76 ± 0.23	3.51 ± 1.41	96.85 ± 0.50
PLANet	2021	91.92 ± 0.34	3.42 ± 0.67	97.41 ± 0.43
Ours	2021	**92.87 ±0.16**	**2.93 ±0.72**	**98.20 ±0.21**

Table 4
Statistical comparison with different state-of-the-art methods on the CAMUS dataset. Numbers format: mean value ± (standard deviation).

Method	Year	DC (%)	HD (mm)	AUC (%)
DFF	2017	90.48 ± 0.34	3.67 ± 1.43	95.90 ± 0.48
Accel	2019	90.81 ± 0.33	3.76 ± 0.72	95.58 ± 0.39
TDNet	2021	90.68 ± 0.23	3.36 ± 0.62	96.72 ± 0.26
Joint-motion	2018	91.98 ± 0.23	3.07 ± 1.53	97.36 ± 0.31
C-LSTM	2018	91.54 ± 0.17	3.34 ± 0.91	97.24 ± 0.21
MFP-Net	2019	92.23 ± 0.29	3.40 ± 0.97	97.28 ± 0.23
Joint-net	2020	91.05 ± 0.27	3.41 ± 0.86	97.14 ± 0.25
TransUNet	2021	91.89 ± 0.38	3.25 ± 1.01	97.39 ± 0.24
PLANet	2021	92.61 ± 0.40	3.10 ± 0.93	97.58 ± 0.23
Ours	2021	**93.79 ±0.26**	**2.86 ±0.66**	**98.52 ±0.23**

icantly reduce the computational cost in echocardiogram segmentation.

In addition, we also have evaluated the impact of different encoders in our semi-supervised segmentation model. Statistical results for the different number of encoders (k) applied on both the EchoNet-Dynamic and CAMUS datasets are as shown in Table 5. From the performance comparison among the different number of frames with k independent encoders in our method, we can clearly observe that we can obtain a better accuracy performance by considering more neighbouring frames in our adaptive semantic calibration and bi-directional semantics fusion modules. However, the running time also distinctly increases with the number for k. To simultaneously guarantee a real-time performance and a relatively better segmentation accuracy of the echocardiography videos, we finally choose $k = 3$ in our experiments.

Moreover, we also have conducted a statistical comparison of left ventricular ejection fraction (LV_{EF}) with different state-of-the-art methods on both the EchoNet-Dynamic and CAMUS datasets. Based on the ground truth and the prediction, we can calculate the correlation coefficient (corr), average deviation (bias), and standard deviation (std) of LV_{EF}, as shown in Table 6. It is also evident from Table 6 that our proposed method can achieve a higher correlation and lower deviation on LV_{EF} than all other methods, which is potentially used to analyze and evaluate cardiac function through echocardiographic segmentation clinically.

Table 6
Statistical comparison of left ventricular ejection fraction (LV_{EF}) with different state-of-the-art methods on the EchoNet-Dynamic and CAMUS datasets.

Method	LV_{EF}		LV_{EF}	
	EchoNet-Dynamic		CAMUS	
	corr	| bias | ± std(%)	corr	| bias | ± std(%)
DFF	0.476	9.2 ± 13.4	0.715	1.4 ± 8.0
Accel	0.593	7.1 ± 12.3	0.749	1.1 ± 8.4
TDNet	0.692	4.6 ± 10.5	0.773	0.9 ± 7.4
Joint-motion	0.761	3.1 ± 9.6	0.823	0.5 ± 5.7
C-LSTM	0.755	3.3 ± 8.9	0.801	0.6 ± 7.3
Joint-net	0.796	2.7 ± 7.2	0.792	0.8 ± 7.9
TransUNet	0.780	2.5 ± 7.6	0.835	0.5 ± 5.6
PLANet	0.826	2.1 ± 6.9	0.858	0.3 ± 6.1
Ours	**0.861**	**1.8 ±7.5**	**0.876**	**0.3 ±5.3**

5. Discussions

Through the above-mentioned ablation studies and comparative experiments, we have observed that our network successfully handles echocardiographic segmentation with real-time performance. Unlike general video semantic segmentation methods, these methods either input the video to the network frame by frame without considering the correlation between time frames, or only repeatedly apply several frames to the same encoder, and simply collect the extracted feature in the time domain maps. Our temporal context-aware feature extraction module can learn different feature representations in the same frame but at different relative time positions. On the other hand, the existing state-of-the-art optical flow-based methods, such as TDNET and ACCEl, cannot resist noise in echocardiograms and relatively irregular and anisotropic heart movements. More importantly, optical flow's Dense pixel-to-pixel mapping also requires a much higher computational cost. Our adaptive Spatiotemporal semantic calibration is based on convolution. Obviously, based on a wider signal bandwidth to describe the left ventricular motion between adjacent frames, our method has stronger resistance to ultrasonic noise and thus improves the accuracy of feature warping. Finally, our method has achieved the latest speed and real-time requirements for clinical applications. By integrating TCE and BSSF modules in our method, we can not only minimize the impact of speckle-noise to improve accuracy performance but also significantly reduce the computational cost of echocardiographic segmentation.

5.1. Speed analysis

This work proposes a segmentation method of the left ventricle endocardium based on the consistency of temporal sequence in

Table 5
Performance comparison of the different number of frames in our temporal context-aware encoder applied on the datasets, where $k = 1$ indicates using a shared encoder.

k	EchoNet-Dynamic			CAMUS			Time (ms/f)
	DC (%)	HD (mm)	AUC (%)	DC (%)	HD (mm)	AUC (%)	
1	90.67 ± 0.22	3.92 ± 0.97	96.13 ± 0.19	91.81 ± 0.33	3.87 ± 1.21	96.68 ± 0.09	12
3	92.87 ± 0.16	2.93 ± 0.72	98.20 ± 0.21	93.79 ± 0.26	2.86 ± 0.66	98.52 ± 0.23	32
5	93.04 ± 0.17	2.90 ± 0.74	98.16 ± 0.23	94.07 ± 0.41	2.81 ± 0.79	98.66 ± 0.22	51
7	93.13 ± 0.16	2.86 ± 0.62	97.94 ± 0.36	94.21 ± 0.27	2.77 ± 0.93	98.38 ± 0.25	71
9	93.19 ± 0.21	2.84 ± 0.81	98.17 ± 0.22	94.33 ± 0.34	2.75 ± 0.72	98.55 ± 0.27	91
11	93.23 ± 0.15	2.83 ± 0.73	98.45 ± 0.17	94.39 ± 0.41	2.73 ± 0.70	98.63 ± 0.21	110

Table 7
Efficiency comparison with the state-of-the-art methods.

Method	Year	Flops (G)	Params (M)	Speed (ms/f)
DFF	2017	200.556	27.084	167
Accel	2019	204.991	61.033	264
TDNet	2021	157.925	23.163	141
Joint-motion	2018	237.592	**17.315**	154
C-LSTM	2018	307.293	21.651	529
Joint-net	2020	108.318	117.266	62
TransUNet	2021	98.518	93.192	46
PLANet	2021	74.954	20.746	34
Ours	2021	**56.359**	74.798	**32**

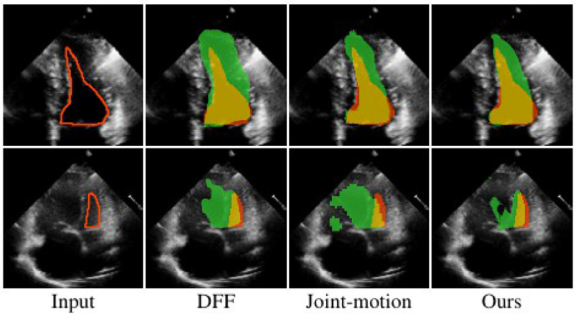

Fig. 7. Failure cases. DFF (Zhu et al., 2017) and Joint-motion (Qin et al., 2018) are two most representative networks for natural and medical video segmentation. Red, green, and yellow regions represent the ground truth, perdition, and the overlapping region between them, respectively.

echocardiography. The main motivation of this approach is to design an adaptive calibration mechanism to utilize the spatiotemporal coherence between neighbouring frames to deal with the main challenges of background speckle noise in ultrasound video segmentation. Although the experimental results show that our proposed method has achieved satisfactory segmentation results, recent studies may suggest that simply increasing the complexity and calculation time of the network can generally lead to better performance. However, we should ensure low latency and real-time performance of the network in ultrasound video segmentation. Therefore, we have further compared our method with other state-of-the-art methods by estimating the number of parameters, calculations, and segmentation speed of the network. As shown in Table 7, compared with other competitors, although our method has increased the number of parameters, it can still achieve the least amount of calculation and the fastest segmentation performance. Compared with the optical flow-based method, our method is a relatively lightweight and efficient network by extracting more semantic information and generating more representative features based on the TCE and BSSF modules.

5.2. Limitations

Although ablation studies and comparisons have proven the advantages of our proposed method, our method still has limitations. Similar to most existing methods, our method still may fail in segmenting the echocardiogram video with extremely irregular cardiac motion or extremely low contrast between the ventricle and the surrounding tissues, as shown in Fig. 7. However, as our method can handle a lot of challenging echocardiogram video segmentation well, our method still potentially becomes a useful tool for clinical echocardiogram applications.

6. Conclusion

In this work, we present a novel model for echocardiography video segmentation, which is of great significance in clinical practice and yet a challenge considering the low image quality of the video and sparse annotation. Based on a novel adaptive spatiotemporal semantic calibration method to align the feature maps of consecutive frames, we can determine spatiotemporal correspondences based on feature maps instead of pixels, thereby mitigating the adverse effects of speckle noise in the calibration. By further learning the importance of each feature map of neighbouring frames to the current frame from the temporal perspective, we can distinctively harness the temporal information to tackle the irregular and anisotropic motions. Extensive experimental results are conducted on two public echocardiography video datasets (EchoNet-Dynamic and CAMUS), demonstrating superior performance over current state-of-the-art methods on echocardiographic video segmentation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Huisi Wu: Conceptualization, Funding acquisition, Project administration, Methodology, Writing – original draft. **Jiasheng Liu:** Conceptualization, Methodology, Writing – original draft. **Fangyan Xiao:** Conceptualization, Methodology, Writing – original draft. **Zhenkun Wen:** Supervision, Funding acquisition, Project administration. **Lan Cheng:** Conceptualization, Methodology, Writing – original draft. **Jing Qin:** Conceptualization, Funding acquisition, Project administration, Methodology, Writing – original draft.

Acknowledgments

This work was supported partly by National Natural Science Foundation of China (No. 61973221), Natural Science Foundation of Guangdong Province, China (Nos. 2018A030313381 and 2019A1515011165), the COVID-19 Prevention Project of Guangdong Province, China (No. 2020KZDZX1174), the Major Project of the New Generation of Artificial Intelligence (No. 2018AAA0102900), and the Hong Kong Research Grant Council under General Research Fund Scheme (Project no. 15205919).

References

Ahn, S.S., Ta, K., Thorn, S., Langdon, J., Sinusas, A.J., Duncan, J.S., 2021. Multi-frame attention network for left ventricle segmentation in 3D echocardiography. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 348–357.

Ali, R., Sheng, B., Li, P., Chen, Y., Li, H., Yang, P., Jung, Y., Kim, J., Chen, C.P., 2020. Optic disk and cup segmentation through fuzzy broad learning system for glaucoma screening. IEEE Trans. Ind. Inf. 17 (4), 2476–2487.

Bai, W., Suzuki, H., Qin, C., Tarroni, G., Oktay, O., Matthews, P.M., Rueckert, D., 2018. Recurrent neural networks for aortic image sequence segmentation with sparse annotations. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 586–594.

Baskaran, L., Maliakal, G., Al Aref, S.J., Singh, G., Xu, Z., Michalak, K., Dolan, K., Gianni, U., van Rosendael, A., van den Hoogen, I., et al., 2020. Identification and quantification of cardiovascular structures from CCTA: an end-to-end, rapid, pixel-wise, deep-learning method. Cardiovasc. Imaging 13 (5), 1163–1171.

Chen, C., Qin, C., Qiu, H., Tarroni, G., Duan, J., Bai, W., Rueckert, D., 2020. Deep learning for cardiac image segmentation: areview. Front. Cardiovas. Med. 7, 25.

Chen, C., Qin, C., Qiu, H., Tarroni, G., Duan, J., Bai, W., Rueckert, D., 2020. Deep learning for cardiac image segmentation: areview. Front. Cardiovas. Med. 7, 25.

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y., 2021. TransUNet: transformers make strong encoders for medical image segmentation. CoRR. abs/2102.04306.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2017. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell. 40 (4), 834–848.

Chen, X., Li, Z., Yuan, Y., Yu, G., Shen, J., Qi, D., 2020. State-aware tracker for real-time video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9384–9393.

Chen, Y., Zhang, X., Haggerty, C.M., Stough, J.V., 2021. Assessing the generalizability of temporally coherent echocardiography video segmentation. In: Medical Imaging 2021: Image Processing, vol. 11596. International Society for Optics and Photonics, p. 115961O.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B., 2016. The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3213–3223.

Ding, M., Wang, Z., Zhou, B., Shi, J., Lu, Z., Luo, P., 2020. Every frame counts: joint learning of video segmentation and optical flow. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 10713–10720.

Eelbode, T., Bertels, J., Berman, M., Vandermeulen, D., Maes, F., Bisschops, R., Blaschko, M.B., 2020. Optimization for medical image segmentation: theory and practice when evaluating with dice score or Jaccard index. IEEE Trans. Med. Imaging 39 (11), 3679–3690.

Gadde, R., Jampani, V., Gehler, P.V., 2017. Semantic video CNNs through representation warping. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4453–4462.

Hou, R., Ma, B., Chang, H., Gu, X., Shan, S., Chen, X., 2020. IAUnet: global context-aware feature learning for person reidentification. IEEE Trans. Neural Netw. Learn. Syst..

Hu, P., Caba, F., Wang, O., Lin, Z., Sclaroff, S., Perazzi, F., 2020. Temporally distributed networks for fast video semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8818–8827.

Hu, Y., Guo, L., Lei, B., Mao, M., Jin, Z., Elazab, A., Xia, B., Wang, T., 2019. 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, pp. 6561–6564.

Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T., 2017. FlowNet 2.0: evolution of optical flow estimation with deep networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2462–2470.

Jafari, M.H., Girgis, H., Liao, Z., Behnami, D., Abdi, A., Vaseli, H., Luong, C., Rohling, R., Gin, K., Tsang, T., et al., 2018. A unified framework integrating recurrent fully-convolutional networks and optical flow for segmentation of the left ventricle in echocardiography data. In: Deep Learning in Medical Image Analysis - and - Multimodal Learning for Clinical Decision Support - 4th International Workshop, pp. 29–37.

Jain, S., Wang, X., Gonzalez, J.E., 2019. Accel: a corrective fusion network for efficient semantic segmentation on video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8866–8875.

Kroeger, T., Timofte, R., Dai, D., Van Gool, L., 2016. Fast optical flow using dense inverse search. In: European Conference on Computer Vision, pp. 471–488.

Kroeger, T., Timofte, R., Dai, D., Van Gool, L., 2016. Fast optical flow using dense inverse search. In: European Conference on Computer Vision. Springer, pp. 471–488.

Leclerc, S., Smistad, E., Grenier, T., Lartizien, C., Ostvik, A., Cervenansky, F., Espinosa, F., Espeland, T., Berg, E.A.R., Jodoin, P.-M., et al., 2019. RU-Net: a refining segmentation network for 2D echocardiography. In: IEEE International Ultrasonics Symposium, pp. 1160–1163.

Leclerc, S., Smistad, E., Pedrosa, J., Østvik, A., Cervenansky, F., Espinosa, F., Espeland, T., Berg, E.A.R., Jodoin, P.-M., Grenier, T., et al., 2019. Deep learning for segmentation using an open large-scale dataset in 2D echocardiography. IEEE Trans. Med. Imaging 38 (9), 2198–2210.

Li, K., Wang, S., Yu, L., Heng, P.-A., 2020. Dual-teacher++: exploiting intra-domain and inter-domain knowledge with reliable transfer for cardiac segmentation. IEEE Trans. Med. Imaging.

Li, M., Zhang, W., Yang, G., Wang, C., Zhang, H., Liu, H., Zheng, W., Li, S., 2019. Recurrent aggregation learning for multi-view echocardiographic sequences segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 678–686.

Li, Y., Shi, J., Lin, D., 2018. Low-latency video semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5997–6005.

Lin, C.-C., Hung, Y., Feris, R., He, L., 2020. Video instance segmentation tracking with a modified VAE architecture. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13147–13157.

Lin, H., Qi, X., Jia, J., 2019. AGSS-VOS: attention guided single-shot video object segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3949–3957.

Liu, F., Wang, K., Liu, D., Yang, X., Tian, J., 2021. Deep pyramid local attention neural network for cardiac structure segmentation in two-dimensional echocardiography. Med. Image Anal. 67, 101873.

Liu, Q., Yu, L., Luo, L., Dou, Q., Heng, P.A., 2020. Semi-supervised medical image classification with relation-driven self-ensembling model. IEEE Trans. Med. Imaging 39 (11), 3429–3440.

Liu, R., Liua, M., Sheng, B., Li, H., Li, P., Song, H., Zhang, P., Jiang, L., Shen, D., 2021. NHBS-Net: a feature fusion attention network for ultrasound neonatal hip bone segmentation. IEEE Trans. Med. Imaging.

Lu, X., Wang, W., Ma, C., Shen, J., Shao, L., Porikli, F., 2019. See more, know more: Unsupervised video object segmentation with co-attention siamese networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3623–3632.

Moradi, S., Oghli, M.G., Alizadehasl, A., Shiri, I., Oveisi, N., Oveisi, M., Maleki, M., Dhooge, J., 2019. MFP-Unet: a novel deep learning based approach for left ventricle segmentation in echocardiography. Physica Med. 67, 58–69.

Moradi, S., Oghli, M.G., Alizadehasl, A., Shiri, I., Oveisi, N., Oveisi, M., Maleki, M., Dhooge, J., 2019. MFP-Unet: a novel deep learning based approach for left ventricle segmentation in echocardiography. Physica Med. 67, 58–69.

Nazir, A., Cheema, M.N., Sheng, B., Li, H., Li, P., Yang, P., Jung, Y., Qin, J., Kim, J., Feng, D.D., 2020. OFF-eNET: an optimally fused fully end-to-end network for automatic dense volumetric 3D intracranial blood vessels segmentation. IEEE Trans. Image Process. 29, 7192–7202.

Nilsson, D., Sminchisescu, C., 2018. Semantic video segmentation by gated recurrent flow propagation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6819–6828.

Noble, J.A., Boukerroui, D., 2006. Ultrasound image segmentation: a survey. IEEE Trans. Med. Imaging 25 (8), 987–1010.

Ouyang, D., He, B., Ghorbani, A., Yuan, N., Ebinger, J., Langlotz, C.P., Heidenreich, P.A., Harrington, R.A., Liang, D.H., Ashley, E.A., et al., 2020. Video-based ai for beat-to-beat assessment of cardiac function. Nature 580 (7802), 252–256.

Painchaud, N., Skandarani, Y., Judge, T., Bernard, O., Lalande, A., Jodoin, P.-M., 2020. Cardiac segmentation with strong anatomical guarantees. IEEE Trans. Med. Imaging 39 (11), 3703–3713.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A., 2017. Automatic differentiation in pytorch.

Pedrosa, J., Queirós, S., Bernard, O., Engvall, J., Edvardsen, T., Nagel, E., D'hooge, J., 2017. Fast and fully automatic left ventricular segmentation and tracking in echocardiography using shape-based b-spline explicit active surfaces. IEEE Trans. Med. Imaging 36 (11), 2287–2296.

Pfeuffer, A., Dietmayer, K., 2020. Robust semantic segmentation in adverse weather conditions by means of fast video-sequence segmentation. In: 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC). IEEE, pp. 1–6.

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

Shelhamer, E., Rakelly, K., Hoffman, J., Darrell, T., 2016. Clockwork convnets for video semantic segmentation. In: European Conference on Computer Vision. Springer, pp. 852–868.

Sheng, B., Li, P., Mo, S., Li, H., Hou, X., Wu, Q., Qin, J., Fang, R., Feng, D.D., 2018. Retinal vessel segmentation using minimum spanning superpixel tree detector. IEEE Trans. Cybern. 49 (7), 2707–2719.

Smistad, E., Østvik, A., Salte, I.M., Melichova, D., Nguyen, T.M., Haugaa, K., Brunvand, H., Edvardsen, T., Leclerc, S., Bernard, O., et al., 2020. Real-time automatic ejection fraction and foreshortening detection using deep learning. IEEE Trans. Ultrason. Ferroelectr. Freq. Control 67 (12), 2595–2604.

Smistad, E., Østvik, A., et al., 2017. 2D left ventricle segmentation using deep learning. In: 2017 IEEE International Ultrasonics Symposium, pp. 1–4.

Sun, D., Yang, X., Liu, M.-Y., Kautz, J., 2018. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8934–8943.

Ta, K., Ahn, S.S., Stendahl, J.C., Sinusas, A.J., Duncan, J.S., 2020. A semi-supervised joint network for simultaneous left ventricular motion tracking and segmentation in 4D echocardiography. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 468–477.

Tarvainen, A., Valpola, H., 2017. Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results. In: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4–9, 2017, Long Beach, CA, USA, pp. 1195–1204.

Veni, G., Moradi, M., Bulu, H., Narayan, G., Syeda-Mahmood, T., 2018. Echocardiography segmentation based on a shape-guided deformable model driven by a fully convolutional network prior. In: 2018 IEEE 15th International Symposium on Biomedical Imaging, pp. 898–902.

Wang, B., Li, L., Nakashima, Y., Kawasaki, R., Nagahara, H., Yagi, Y., 2021. Noisy-LSTM: improving temporal awareness for video semantic segmentation. IEEE Access 9, 46810–46820.

Wang, G., Liu, X., Li, C., Xu, Z., Ruan, J., Zhu, H., Meng, T., Li, K., Huang, N., Zhang, S., 2020. A noise-robust framework for automatic segmentation of COVID-19 pneumonia lesions from ct images. IEEE Trans. Med. Imaging 39 (8), 2653–2663.

Wang, H., Wang, W., Liu, J., 2021b. Temporal memory attention for video semantic segmentation. arXiv preprint arXiv:2102.08643.

Wang, Z., Xu, J., Liu, L., Zhu, F., Shao, L., 2019. RANet: ranking attention network for fast video object segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3978–3987.

Wei, H., Cao, H., Cao, Y., Zhou, Y., Xue, W., Ni, D., Li, S., 2020. Temporal-consistent segmentation of echocardiography with co-learning from appearance and shape. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 623–632.

Wu, H., Lu, X., Lei, B., Wen, Z., 2021. Automated left ventricular segmentation from cardiac magnetic resonance images via adversarial learning with multi-stage pose estimation network and co-discriminator. Med. Image Anal. 68, 101891.

Wu, H., Pan, J., Li, Z., Wen, Z., Qin, J., 2020. Automated skin lesion segmentation via an adaptive dual attention module. IEEE Trans. Med. Imaging 40 (1), 357–370.

Wu, H., Wang, W., Zhong, J., Lei, B., Wen, Z., Qin, J., 2021. SCS-Net: a scale and context sensitive network for retinal vessel segmentation. Med. Image Anal. 70, 102025.

Zhang, L., Lin, Z., Zhang, J., Lu, H., He, Y., 2019. Fast video object segmentation via dynamic targeting network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5582–5591.

Zhou, H., Qi, L., Wan, Z., Huang, H., Yang, X., 2020. RGB-D co-attention network for semantic segmentation. In: Proceedings of the Asian Conference on Computer Vision.

Zhu, X., Xiong, Y., Dai, J., Yuan, L., Wei, Y., 2017. Deep feature flow for video recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2349–2358.