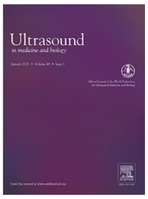
Ultrasound in Medicine & Biology 50 (2024) 1945−1954

Contents lists available at ScienceDirect

Ultrasound in Medicine & Biology

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



Original Contribution

SimLVSeg: Simplifying Left Ventricular Segmentation in 2-D + Time Echocardiograms With Self- and Weakly Supervised Learning

Fadillah Maani *, Asim Ukaye, Nada Saadi, Numan Saeed, Mohammad Yaqub

Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates

ARTICLE INFO

Keywords:
Left ventricular segmentation
Sparse video segmentation
Self-supervision
Temporal masking
3-D segmentation
Super image

ABSTRACT

Objective: Achieving reliable automatic left ventricle (LV) segmentation from echocardiograms is challenging due to the inherent sparsity of annotations in the dataset, as clinicians typically only annotate two specific frames for diagnostic purposes. Here we aim to address this challenge by introducing simplified LV segmentation (SimLVSeg), a novel paradigm that enables video-based networks for consistent LV segmentation from sparsely annotated echocardiogram videos.

Methods: SimLVSeg consists of two training stages: (i) self-supervised pre-training with temporal masking, which involves pre-training a video segmentation network by capturing the cyclic patterns of echocardiograms from largely unannotated echocardiogram frames, and (ii) weakly supervised learning tailored for LV segmentation from sparse annotations.

Results: We extensively evaluated SimLVSeg using EchoNet-Dynamic, the largest echocardiography dataset. SimLVSeg outperformed state-of-the-art solutions by achieving a 93.32% (95% confidence interval: 93.21 −93.43%) dice score while being more efficient. We further conducted an out-of-distribution test to showcase SimLVSeg's generalizability on distribution shifts (CAM US dataset).

Conclusion: Our findings show that SimLVSeg exhibits excellent performance on LV segmentation with a relatively cheaper computational cost. This suggests that adopting video-based networks for LV segmentation is a promising research direction to achieve reliable LV segmentation. Our code is publicly available at https://github.com/BioMedIA-MBZUAI/SimLVSeg.

Introduction

Echocardiograms are a crucial modality in cardiovascular imaging due to their safety, availability and high temporal resolution [1]. In clinical practice, echocardiogram information is used to diagnose heart conditions and understand pre-operative risks in patients with cardiovascular diseases [2]. Through heartbeat sequences in echocardiogram videos, clinicians measure ejection fraction (EF) to assess the heart's capability to supply adequate oxygenated blood. The EF reflects the percentage of blood the heart can pump out of the left ventricle (LV), which is calculated as (EDV − ESV)/EDV, using the LV volume in the end-diastole (ED) phase and end-systole (ES) phase [3]. ED refers to the phase where the heart is maximally filled with blood just before contraction, while the ES phase happens immediately after the contraction where the volume of the heart chambers is in its minimum stage. By accurately segmenting the heart structures, especially on ED and ES frames, clinicians can assess the heart's condition, detect any symptoms, determine the appropriate treatment approach and monitor the patient's response to therapy [4].

The typical manual workflow of segmenting the LV is as follows [5−7]: (i) a sonographer acquires an echocardiogram video using an ultrasound device and records the patient's heartbeat; (ii) the sonographer then finds ED and ES by locating candidate frames indicated by the recorded heartbeat signal and then verifies them visually with the recorded echocardiogram video; (iii) and ultimately draws some key points to represent the LV structure, as shown in Figure 1. The manual LV segmentation workflow is typically time-consuming and prone to intra- and inter-observer variability [5,8−10]. Variability in image quality and the presence of noise in echocardiograms make LV segmentation more challenging [10,11] as the LV boundaries are sometimes unclear [11]. Hence, sonographers must consider the temporal context to eliminate any ambiguity caused by unclear heart structures in echocardiograms and perfectly segment the LV to achieve accurate results, which unfortunately means adding more burden for sonographers because they must go back and forth between echocardiogram frames to analyze the ambiguous boundaries properly. Automatic LV segmentation can help sonographers to solve this arduous task more efficiently [7,12].

* Corresponding author. Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates.
E-mail address: fadillah.maani@mbzuai.ac.ae (F. Maani).

https://doi.org/10.1016/j.ultrasmedbio.2024.08.023
Received 25 March 2024; Revised 23 July 2024; Accepted 27 August 2024

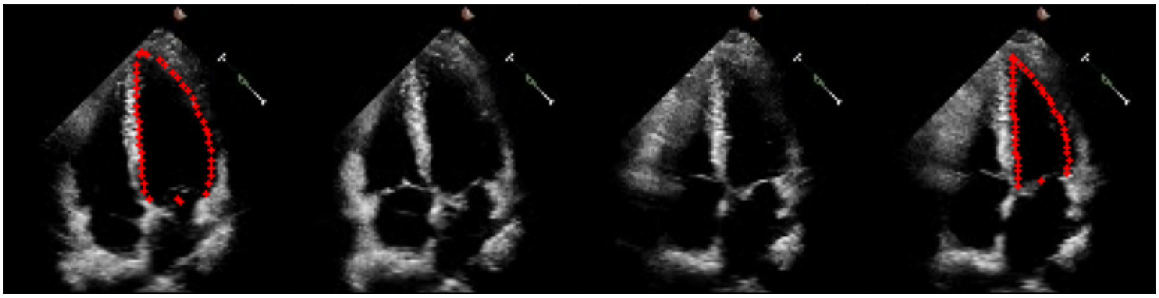


Figure 1. Sequence of an echocardiogram video [7]. The number of frames varies, yet only two are labeled, *i.e.*, the end-diastole (*left-most*) and the end-systole (*right-most*) frame. Annotators draw key points to represent the left ventricular (LV) region. Then, LV segmentation labels are inferred from the given key points.

A wide range of work on medical image segmentation using a supervised deep-learning approach has been presented [13,14]. The problem in echocardiogram segmentation, however, is more challenging as clinicians usually provide only two annotated frames per video—*i.e.*, ED and ES frames—resulting in limited labels for supervision. For instance, in EchoNet-Dynamic [7], the largest publicly available 2-D + time echocardiography dataset, this utilizes less than 1.2% of the available frames when training in a supervised 2-D setting. Consequently, early studies on LV segmentation proposed a frame-by-frame (2-D) image segmentation solution [7,15−18]. In spite of advancements in segmentation techniques, a recent study [19] surprisingly continues to use this approach due to limited annotations for supervision, fine-tuning the segmentation model from Ouyang et al. [7] to segment apical four-chamber and parasternal short-axis views in pediatric patients. These approaches do not capitalize on the periodicity and temporal consistency of the echocardiograms, which may lead to incoherence in the segmentation results from one frame to the next. In the worst-case scenario, this incoherence can lead to ED- and ES-phase detection failure in the fully automatic EF prediction pipeline [20]. This has motivated recent video-based echocardiogram segmentation approaches.

Recent video-based approaches show high temporal consistency and state-of-the-art performance. Li et al. [21] utilized a set of Conv-LSTM layers to ensure spatiotemporal consistency between consecutive frames. However, their recurrent units incurred a high computational cost. Ahn et al. [22] employed a multi-frame attention network to perform 3-D segmentation. However, multi-frame attention reported by Ahn et al. [22] similarly had computational cost correlated with the number of frames, and they were therefore limited to using five frames. Sirjani et al. [23] used the starting frame of a cardiac cycle and its segmentation label to guide the segmentation of a target frame. Nevertheless, this method did not facilitate automatic segmentation, as the user had to annotate the reference frame during inference. Wu et al. [24] demonstrated the effectiveness of semi-supervision using mean teacher networks and spatiotemporal fusion on segmentation; unfortunately, they limited the temporal context to three frames to obtain optimum performance−compute trade-off. Wei et al. [25] proposed two-stage training to enforce temporal consistency on a 3-D U-Net by leveraging an echocardiogram ED and ES sequence constraint. They leveraged a constraint in their training pipeline where the segmented area changed monotonically as the first input frame was ED and the last frame was ES in the same (one) heartbeat cycle, thus limiting the usage of vastly unannotated frames in other cycles. Chen et al. [26] slightly extended the work of Wei et al. [25] by implementing additional data augmentation proposed by Stough et al. [27]. Painchaud et al. [28] improved the average segmentation performance by enforcing temporal smoothness as a post-processing step on video segmentation outputs. Recently, Yang et al. [29] utilized a 2-D segmentation network and proposed a temporal consistency loss by contrasting frames from the same video as positive pairs and frames from different videos as negative pairs. However, their method enforced temporal consistency at the representation level rather than at the segmentation output level, potentially limiting the effectiveness of enhancing temporal consistency in the final output.

Investigating an alternative research direction, recent work has adopted a self-supervised learning (SSL) technique to effectively utilize unannotated echocardiogram frames. Dezaki et al. [30] introduced a self-supervision method that leveraged spatiotemporal patterns and interdependencies within and between echocardiographic sequences to develop a rich feature-embedding space for temporal synchronization. Similarly, Dai et al. [31] developed a cyclical self-supervision method that leveraged the repetitive pattern of the heartbeat to ensure feature similarity. Saeed et al. [32] used contrastive pre-training to provide self-supervision on echocardiograms. However, these self-supervised methods are tailored for 2-D networks, potentially resulting in low temporal consistency for segmentation. Recent studies in the natural domain, such as those by Feichtenhofer et al. [33] and Tong et al. [34], adopted masked autoencoders for self-supervised pre-training to video networks, enabling accelerated training and showing promising results in action recognition from natural videos.

The aforementioned works perform the LV segmentation from echocardiogram videos either by (i) analyzing frames independently with simple 2-D deep-learning models, or (ii) performing 2-D + time analysis and developing models using complex training schemes. In our proposed method, while achieving state-of-the-art performance, we aimed to mimic clinical assessment where doctors assess multiple frames concurrently in a simplified approach. Thus, we introduced simplified LV segmentation (SimLVSeg), a novel training framework that enables video-based networks for LV segmentation resulting in enhanced performance and higher temporal consistency. SimLVSeg is inspired by the methods presented by Tong et al. [34] and Cicek et al. [35]. We devised a specific design implementation to address the primary issue in LV segmentation. SimLVSeg consists of two training stages: (i) self-supervised pre-training with temporal masking and (ii) weakly supervised learning for LV segmentation, specifically designed to address the challenge of sparsely annotated (labeled) echocardiogram videos. The self-supervised pre-training stage enables a video network to learn the periodic nature of echocardiograms and cardiac patterns, ultimately providing a robust model initialization. Subsequently, the weakly supervised learning stage allows the model to effectively learn LV segmentation from sparse annotations through end-to-end learning. Our main contributions were as follows:

- We introduced a novel paradigm of performing LV segmentation with SimLVSeg. We showed that it is feasible to develop video-based segmentation networks for LV despite the nature of sparsely annotated echocardiogram data. These networks effectively leverage spatial and temporal analysis, ensuring consistency across video frames. SimLVSeg is simple yet effective, opening new research directions for efficient and reliable LV segmentation empowered by video-based segmentation networks.
- We demonstrated how SimLVSeg outperforms the state of the art in LV segmentation on EchoNet-Dynamic, the largest 2-D + time

echocardiography dataset, in terms of performance and efficiency through extensive ablation studies.

- We showed SimLVSeg's compatibility with two types of video segmentation networks: 2-D super image [36,37] and 3-D segmentation networks with various encoder backbones. This indicates that the excellent performance can be attributed to the SimLVSeg design rather than the selection of underlying network architectures.

Methodology

Analyzing both spatial and temporal features in echocardiography is crucial as the heart in 2-D + time echocardiography is dynamic. Certain structures may not be visible in a single echocardiogram frame but become apparent in other frames. However, training a video-based segmentation network for LV segmentation is challenging due to sparse labeling. To address this, we proposed SimLVSeg, which is illustrated in Figure 2. SimLVSeg is composed of a self-supervised temporal masking approach that leverages vastly unannotated echocardiogram frames to provide better network initialization for the downstream LV segmentation task by learning the periodic nature of echocardiograms, and weakly supervised training that allows a video-based segmentation network to learn the LV segmentation from sparsely annotated (labeled) echocardiogram videos without any heartbeat cycle constraint. The network utilizes unannotated frames for a pre-training stage and learns from annotated frames in a weakly supervised manner. The performance of the proposed method was evaluated with 3-D segmentation and 2-D super image (SI) segmentation [36,37] approach, as depicted in Figure 3. The details are described below.

Self-supervised temporal masking

In the EchoNet-Dynamic [7] dataset most of the frames were unannotated, thus the ability to perform supervised training was limited. To benefit from the vast amount of unlabeled frames, we implemented a self-supervised temporal masking algorithm to pre-train our model. As depicted in Figure 2, a clip of an echocardiogram video was retrieved and a portion of the frames was randomly selected and masked by setting the pixel values to 0. This random selection for masking was designed to prevent the model from relying on specific frames. The model was then pre-trained to reconstruct the masked clip. Through this process, the model learned valuable latent information from the periodic nature of echocardiograms, *e.g.*, the embedded temporal pattern or cardiac rhythm, that benefited the downstream LV segmentation task.

More formally, suppose V is an echocardiogram video with H × W frame size. From V, we sampled a clip $v \in \mathbb{R}^{H \times W \times F \times 3}$consisting of F number of consecutive frames with a stride or sampling period of T. Then, we provided a masked clip $v_m \in \mathbb{R}^{H \times W \times F \times 3}$ by randomly choosing F_m number of frames ($F_m < F$) from v and adjusting their pixel values to 0. A video network G_Ψ with a set of parameters $_\Psi$ was then pre-trained to reconstruct v from v_m. The network G_Ψ was optimized by minimizing the following objective, as shown in eqn (1):

$$L_{rec} = \frac{1}{N}\sum_{n-1}^{N} MSE\left(v^n, G_\Psi({v_m}^n)\right) \tag{1}$$

where N is the batch size.

Weakly supervised LV segmentation with sparse annotation

Sparsely annotated echocardiogram videos make LV segmentation challenging as training a video segmentation model on EchoNet-Dynamic is not trivial. To tackle this issue, and inspired by Cicek et al. [35], we proposed a training strategy to develop a video segmentation network specifically for LV. As illustrated in Figure 2, the network took in F number of frames and segmented the LV on each frame. Then, the loss was calculated and back-propagated only based on the prediction of frames having a segmentation label.

More formally, G_Ψ was the pre-trained video segmentation network, which took in an input echocardiogram clip $v \in \mathbb{R}^{H \times W \times F \times C}$ and predicted LV segmentation $\hat{y} = \{\hat{y}_1, \hat{y}_2, \cdots \hat{y}_F\}, \hat{y}_i \in \mathbb{R}^{H \times W}$, where F, C and $H \times W$ were the number of frames, the number of channels (which is 3) and the frame size, respectively. Additionally, $\boldsymbol{y} = \{y_1, y_2, \cdots y_F\}, y_i \in \mathbb{R}^{H \times W}$ denoted the sparse segmentation label of the input clip where most y_i were empty. Thus, we constructed $\boldsymbol{y}$ for every sample by using the following rule (eqn [2]):

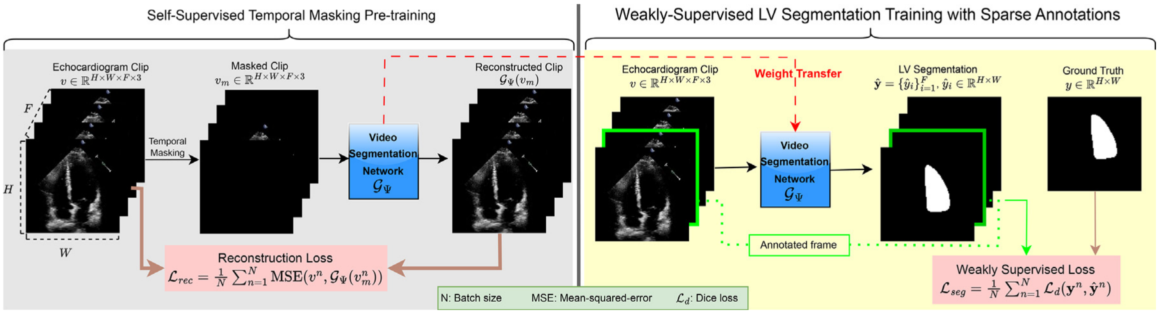

Self-Supervised Temporal Masking Pre-training
Weakly-Supervised LV Segmentation Training with Sparse Annotations
Echocardiogram Clip
$v \in \mathbb{R}^{H \times W \times F \times 3}$
F
H
W
Temporal Masking
Masked Clip
$v_m \in \mathbb{R}^{H \times W \times F \times 3}$
Video Segmentation Network
$\mathcal{G}_\Psi$
Reconstructed Clip
$\mathcal{G}_\Psi(v_m)$
Reconstruction Loss
$\mathcal{L}_{rec} = \frac{1}{N}\sum_{n=1}^{N} \mathrm{MSE}(v^n, \mathcal{G}_\Psi(v_m^n))$
Echocardiogram Clip
$v \in \mathbb{R}^{H \times W \times F \times 3}$
Weight Transfer
Video Segmentation Network
$\mathcal{G}_\Psi$
LV Segmentation
$\hat{\mathbf{y}} = \{\hat{y}_i\}_{i=1}^{F}, \hat{y}_i \in \mathbb{R}^{H \times W}$
Ground Truth
$y \in \mathbb{R}^{H \times W}$
Annotated frame
Weakly Supervised Loss
$\mathcal{L}_{seg} = \frac{1}{N}\sum_{n=1}^{N} \mathcal{L}_d(\mathbf{y}^n, \hat{\mathbf{y}}^n)$
N: Batch size
MSE: Mean-squared-error
$\mathcal{L}_d$: Dice loss

Figure 2. Illustration of SimLVSeg. A video segmentation network is developed to segment the left ventricle (LV) on every input echocardiogram frame. The network is pre-trained using a self-supervised temporal masking method, which is then fine-tuned on the LV segmentation task with sparse annotations.

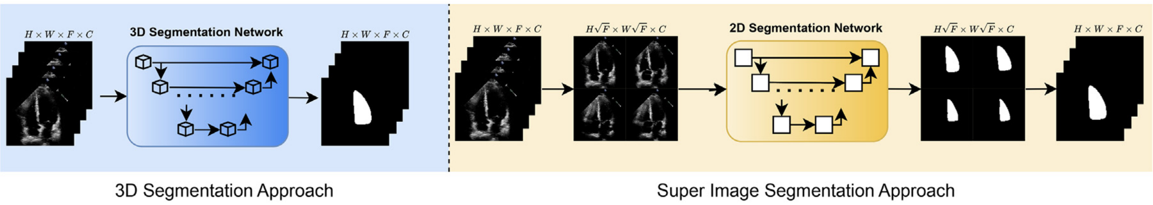

$H \times W \times F \times C$
3D Segmentation Network
$H \times W \times F \times C$
3D Segmentation Approach
$H \times W \times F \times C$
$H\sqrt{F} \times W\sqrt{F} \times C$
2D Segmentation Network
$H\sqrt{F} \times W\sqrt{F} \times C$
$H \times W \times F \times C$
Super Image Segmentation Approach

Figure 3. 3-D versus 2-D super image segmentation approach. The first approach utilizes a 3-D segmentation network, while the second rearranges the echocardiogram clip as a super image and then utilizes a 2-D network.

$$y_i = \begin{cases} y_i & \text{if } i-\text{th frame if } \textit{labeled} \\ \phi & \text{otherwise} \end{cases} \tag{2}$$

Thus, the total dice loss L_d for every sample, n, could be formulated as (eqn [3]):

$$L_d(\boldsymbol{y}^n, \widehat{\boldsymbol{y}}^n) = \sum_{i=1}^{F} \ell_d(y_i^n, \widehat{y}_i^n) = \underbrace{\sum_{j \in F_l^n} \ell_d(y_j^n, \widehat{y}_j^n)}_{\text{labeled (annotated) frames}} + \underbrace{\sum_{k \in \{1,\ldots,F\} \setminus F_l^n} \ell_d(y_k^n, \widehat{y}_k^n)}_{\text{unlabeled frames}} \tag{3}$$

where l_d was the frame-wise dice loss and F_l^n was the set of indices of labeled frames for the n-th sample. The gradient of L_d with respect to a parameter $\psi \in \Psi$ was given by eqn (4):

$$\frac{\partial \mathcal{L}_d}{\partial \psi}(\boldsymbol{y}^n, \hat{\boldsymbol{y}}^n) = \sum_{j \in \mathcal{F}_l^n} \frac{\partial \ell_d}{\partial \psi}\left(y_j^n, \hat{y}_j^n\right) + \sum_{k \in \{1,\ldots,F\} \setminus \mathcal{F}_l^n} \cancelto{0}{\frac{\partial \ell_d}{\partial \psi}\left(y_k^n, \hat{y}_k^n\right)} \tag{4}$$

where $\frac{\partial L_d}{\partial \psi}(y_k^n, \widehat{y}_k^n)$ could be simply set to zero because the k-th frame was unlabeled, preventing the unlabeled frames from contributing to the gradients. Because eqn (1) $\widehat{y}_j \in G_\Psi(v)$ and eqn (2) G_Ψ typically consisted of shared-weights operators (*e.g.*, convolution and attention), eqn (5) was then applied for all parameters ψ in Ψ as follows:

$$\frac{\partial \ell_d}{\partial \psi}(y_j^n, \widehat{y}_j^n) \in \mathbb{R} \Rightarrow \sum_{j \in F_l^n} \frac{\partial \ell_d}{\partial \psi}(y_j^n, \widehat{y}_j^n) \in \mathbb{R} \Rightarrow \frac{\partial L_d}{\partial \psi}(\boldsymbol{y}_j^n, \widehat{\boldsymbol{y}}_j^n) \in \mathbb{R} \tag{5}$$

Thus, although clip v was partially labeled and gradients did not come from unlabeled frames, this framework could facilitate training for all G_Ψ parameters. Ultimately, the total segmentation loss was given by eqn (6):

$$L_{seg} = \frac{1}{N} \sum_{n=1}^{N} L_d(\boldsymbol{y}^n, \widehat{\boldsymbol{y}}^n) \tag{6}$$

Based on this, we can argue that frames without labels assist the network in extracting spatiotemporal features by providing additional context and information to the network input, allowing the network to better understand dynamic heart behavior and structures. Additionally, it is important to note that this weakly supervised learning approach effectively transforms into a fully supervised learning process if every frame has its corresponding segmentation label. However, this is not the typical setting in real-world practice.

During training, a clip is randomly extracted around an annotated frame from every video with the specified number of frames F and sampling period T, resulting in more variations and acting as a regularizer. In other words, there is only a segmentation mask for one frame on every clip. The clip-extraction process is deterministic during the evaluation step to ensure reproducibility. During evaluation, we sampled two clips from each video: one where the annotated ED frame was at the center and another where the annotated ES frame was at the center. Each clip was extracted with the specified number of frames F and sampling period T, ensuring the respective annotated frame was always positioned at the center. These clips were then passed individually to the model to segment the LV. This deterministic approach ensured that the evaluation process was fair across methods and different values of F and T.

Video segmentation

We aimed to develop a video segmentation network, G_Ψ, capable of segmenting LV from an echocardiogram clip,$v \in \mathbb{R}^{H \times W \times F \times C}$. We considered two segmentation approaches as visualized in Figure 3, *i.e.*, the 3-D segmentation approach and the 2-D SI approach. The 3-D approach considered an echocardiogram clip as a 3-D volume, while the SI approach addressed the video segmentation problem in a 2-D fashion [37]. We describe the details of both approaches below.

3-D segmentation approach

Echocardiogram videos consist of stacked 2-D images. Considering the time axis as the third dimension allows 3-D models to segment the LV on an echocardiogram clip. Thus, 3-D U-Net [35] was utilized as the architecture. As depicted in Figure 4, we used a CNN with residual units [38] as the encoder, which has five stages where the stage outputs are passed to the decoder. A residual unit comprises two Conv2D layers, two instance normalization layers, two PReLU activation functions and a skip connection.

2-D SI approach

An echocardiogram clip, v, was rearranged into a single big image $x \in \mathbb{R}^{\widehat{H} \times \widehat{W} \times C}$, where $\widehat{H}$ and $\widehat{W}$were the height and width of the SI, respectively. Because SI works best with a grid layout [36,37], we set the echocardiogram SI size to be $H\sqrt{F} \times W\sqrt{F}$. Hence, existing techniques for 2-D image analysis could be well utilized to help solve the problem, *e.g.*, state-of-the-art architectures, self-supervised methods and strong pre-trained models.

The 2-D U-Net [13] was used as the main architecture, with UniFormerS [39] as the encoder. We selected UniFormer-S as: (i) it leverages the strong properties of convolution and attention, and (ii) it is the recent state of the art on EchoNet-Dynamic EF estimation [40]. In short, the network consisted of four stages, where the first two stages utilized convolution operators to extract features and the rest implemented multi-head self-attention to learn global contexts. The inductive biases

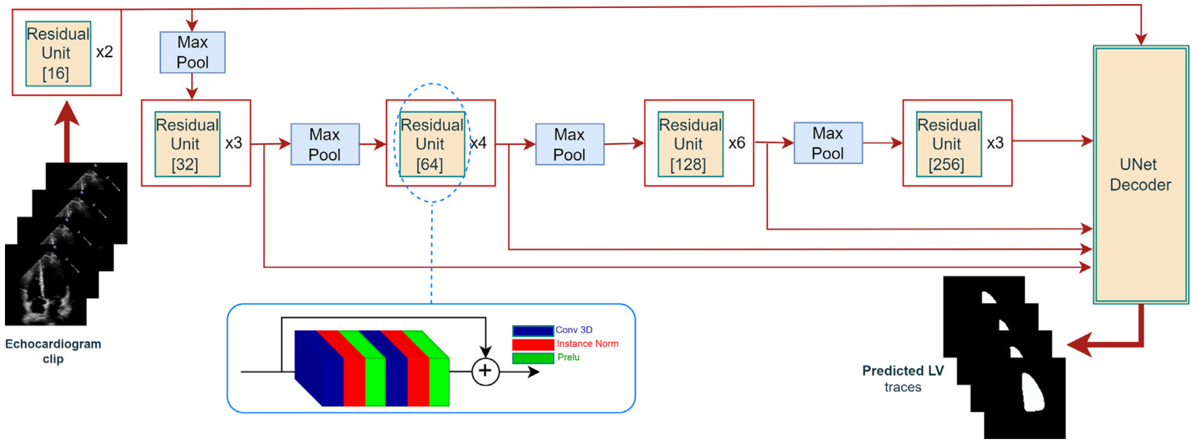


Figure 4. 3-D U-Net architecture. A residual unit [38] consists of convolutional layers, instance norm layers, PReLU and a skip connection. Residual unit [C] denotes a residual unit with C number of feature channels.

of convolution layers allow the model to learn efficiently and multi-head self-attention has a large receptive field that is favorable for SI [36].

Experimental setup

Experiments were mainly performed on EchoNet-Dynamic [7], a large-scale echocardiography dataset, using an NVIDIA RTX 6000 GPU with CUDA 11.7 and PyTorch 1.12. We additionally conducted an out-of-distribution (OOD) test of SimLVSeg on the CAMUS dataset [41], a small echocardiography dataset, broadening the scope of our validation efforts.

Datasets

EchoNet-Dynamic

EchoNet-Dynamic [7] is the largest publicly available dataset of 2-D + time echocardiograms of apical four-chamber views of the human heart. The dataset comprises approximately 10,030 heart echocardiogram videos with a fixed frame size of 112 × 112. Video length varies from 28 to 1002 frames, encompassing multiple heartbeat cycles, yet only two are annotated (ED and ES frames). A sample echocardiogram sequence is given in Figure 1.

To ensure a fair comparison with reported state-of-the-art methods, we adhered strictly to the organizer's provided split, consisting of 7460 training videos, 1288 validation videos and 1276 test videos.

CAMUS

CAMUS [*41*] comprises 500 2-D + time echocardiograms. Each echocardiogram captures a single heartbeat cycle with corresponding dense segmentation labels, i.e., segmentation annotations on all frames. The frame size varies and the video length ranges from 10 to 42 frames, with a median of 20 frames. The ED and ES frames are located at the edges of the echocardiogram video, i.e., the first and last frames. This dataset also includes metadata associated with every echocardiogram scan, such as image quality, patient gender and age.

Implementation details

Main experiments

We conducted our main experiments on the EchoNet-Dynamic dataset. We pre-trained our video segmentation models for 100 epochs with self-supervision. Each echocardiogram video was randomly sampled on every epoch with a specified number of frames (F) and a stride or sampling period (T) to provide more variations. We utilized the AdamW optimizer and set the learning rate to a 3e-4 learning rate and weight decay to 1e-5, with a batch size of 16. A set of augmentations was applied to enrich variation during training, consisting of color jitter (±0.2 for brightness, contrast, saturation and hue), CLAHE, random frame rotation (±20°), padding to 124 × 124 frame size and random cropping to 112 × 112 frame size. The model was then fine-tuned for the LV segmentation task with sparse annotations in a weakly supervised manner for 70 epochs with the same optimizer, batch size and augmentations. Every video was sampled twice on every epoch to accommodate the annotated ED and ES frames. The hyper-parameters were set experimentally.

Out-of-distribution test

We evaluated the OOD performance of SimLVSeg on the CAMUS dataset for LV segmentation. Each echocardiogram was resized to a 112 × 112 frame to align with the EchoNet-Dynamic dataset frame size. For sequences with length shorter than F, we appended zero padding to the temporal axis from the end of the echocardiogram sequence. In contrast, for videos exceeding F frames in length, we uniformly selected F frames across the original sequence. This was achieved by calculating equally spaced indices over the sequence length and rounding these indices to ensure that they corresponded to actual frame numbers. Each selected frame was then extracted to form a new sequence with exactly F frames, ensuring the sequence length matched F. In addition, the OOD test was conducted using medium- and good-quality echocardiogram scans.

Results

Evaluation metrics

We evaluated our method against the baselines on three main areas: (i) segmentation accuracy, (ii) computational cost and (iii) model size.

Segmentation accuracy was measured using the Dice-Srensen coefficient (DSC), which is evaluated as follows (eqn [7]):

$$DSC = \frac{2 * |Y_{pred} \cap Y_{gt}|}{|Y_{pred}| + Y_{gt}|} \quad (7)$$

where Y_{pred} is the predicted pixels for the LV and Y_{gt} is the corresponding ground truth pixels. $|Y_{pred} \cap Y_{gt}|$ represents the area of overlap between the predicted and ground truth pixels. The DSC ranges from 0 to 1; we reported the scores in our experiments as percentages. A higher DSC implies better overlap of the predicted pixels with the ground truth. We reported the overall DSC along with individual scores for the ED and ES frames. We further reported a 95% confidence interval (CI) evaluated through bootstrapping to provide statistical significance of our results.

For the computational cost, we reported the giga floating point operations per second, and the model size was reported as the number of parameters measured in millions.

Comparison with state of the art

SimLVSeg outperformed recent state-of-the-art approaches [7,14,18] on the EchoNet test set, as shown in Table 1 and Figure 5. We compared our method with the approach proposed by the EchoNet dataset publisher [7], the famous nnU-Net [14], which can perform better than specially designed echocardiography networks, as mentioned in [20], and the method achieving the highest DSC on the test set [18]. SimLVSeg with 3-D U-Net (SimLVSeg-3-D) resulted in 93.32% overall DSC, and the SimLVSeg-SI approach showed on-par performance. CI analysis further showed no overlap between the 95% CI of SimLVSeg with other state-of-the-art solutions, indicating that our improvements hold statistical significance over those methods, with a *p* value of less than 0.05. SimLVSeg-3-D was trained with 32 frames sampled consecutively, while SimLVSeg-SI was trained with 16 frames sampled at every fifth frame. The inference run times per frame for nnU-Net [14], SimLVSeg-SI and SimLVSeg-3-D were 3.18 ± 0.04, 0.83 ± 0.02 and 0.79 ± 0.02 ms, respectively, as calculated from 1000 runs. This experiment showed that a video segmentation network trained in a weakly supervised manner is capable of segmenting the LV with a 3.8-times lower computational cost compared with SepXception [18], and it achieved a 4-times faster segmentation speed compared with nnU-Net [14], a widely acclaimed and highly reliable framework for medical imaging segmentation [42]. Additionally, we provided SimLVSeg prediction in challenging echocardiographic scenarios such as foreshortening, poor image quality, effusion and arrhythmia in Figure 6.

Ablation studies

Number of frames and sampling period

The number of frames, F, and the sampling period, T, play important roles [24,40]. A large F allows a network to retrieve rich temporal information while increasing T reduces redundancy between frames. We studied the combination of (F, T) to find the optimum pair, as provided in Figure 7. The (16, 5) combination

Table 1
Dice similarity coefficient on EchoNet-Dynamic test set

Method	DSC (95% CI)			FLOPs	# Parameters
	Overall	ES	ED	(G)	(M)
EchoNet-Dynamic [7]	92.00 (91.87−92.13)	90.68 (90.55−90.86)	92.78 (92.61−92.94)	7.84	39.64
CSS-SemiVideo [31]	92.30 (92.17−92.43)	90.98 (90.75−91.20)	93.14 (92.98−93.29)	7.84	39.64
GraphEcho [29]	−	90.50	93.40	−	−
nnU-Net [14]	92.86 (92.74−92.98)	91.63 (91.43−91.83)	93.62 (93.48−93.76)	2.30	7.37
SepXception [18]	92.90	91.73 (91.54−91.92)	93.64 (93.50−93.78)	4.28	55.83
SimLVSeg-SI	93.31 (93.19−93.43)	92.26 (92.08−92.44)	93.95 (93.81−94.09)	2.17[a]	24.83
SimLVSeg-3D	93.32 (93.21−93.43)	92.29 (92.11−92.47)	93.95 (93.81−94.09)	1.13[a]	18.83

SimLVSeg shows state-of-the-art performance with fewer FLOPs and relatively fewer parameters. fvcore was utilized to count the FLOPs.
DSC, dice similarity coefficient; ED, end diastole; ES, end systole; FLOPs: floating point operations per second; G, Gigaflops; M, Millions of parameters.
[a] Note that we report FLOPs on a per-frame basis.

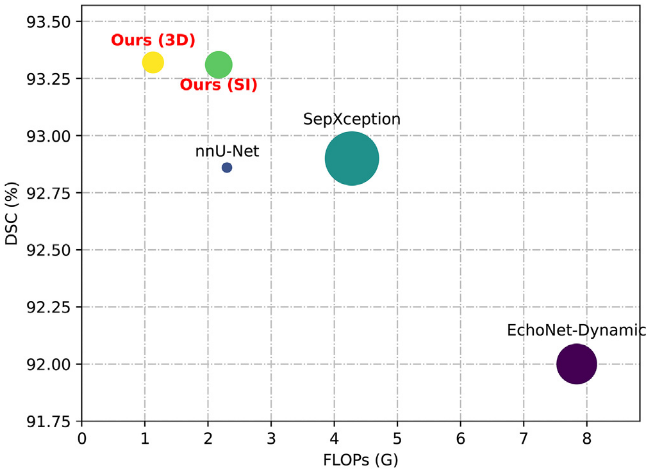


Figure 5. Comparison with other state-of-the-art solutions. Bubble size represents the number of parameters.

resulted in the highest DSC of 93.21% for SI, while (32, 1) gave the best performance for the 3-D approach, resulting in 93.31% DSC. Additionally, all (F, T) pairs resulted in a better performance compared with the recent state of the art [18].

SSL temporal masking

We conducted an ablation study (Fig. 8) to determine the optimum value of the masking ratio and obtain the best results for 60% masking. We found that SSL pre-training helped to maintain better temporal consistency and improve robustness (Fig. 9).

Different backbones

An ablation study was performed on different encoders of the segmentation architecture to see how well our approach adapted to model complexity. We implemented ResNet-18 [43], MobileNet-V3 [44] and ViT-B/16 [45] as SI approach encoders. We also tested a smaller version of 3-D U-Net (Fig. 4), which consists of two residual units on every stage (3-D U-Net-S). As provided in Table 2, the experiment shows that the performance was robust for encoder backbones.

OOD test

We conducted an additional test on the CAMUS dataset to determine if SimLVSeg could generalize well to samples with distributions unseen during training, such as different echocardiogram image contrasts, intensities and original frame aspect ratios. Table 3 presents the average DSC for all frames in the 'Overall' column, as CAMUS provided dense labels or annotations. Additionally, we present the DSC of the middle, ED and ES frames separately. The self-supervised temporal masking pre-training led to improved overall performance with no overlap between

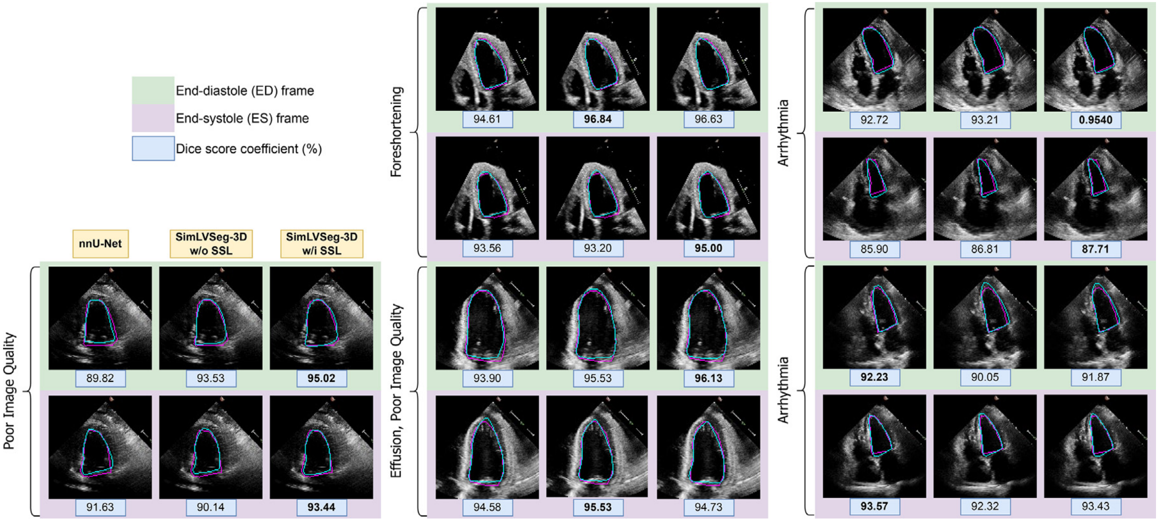


Figure 6. Comparative performance on challenging echocardiograms. This figure demonstrates the efficacy of SimLVSeg compared with nnU-Net in handling complex scenarios such as foreshortening, poor image quality, effusion and arrhythmia.

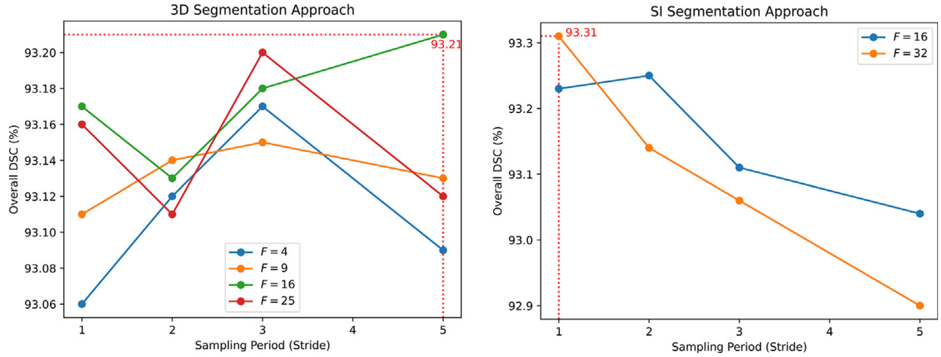


Figure 7. Impact of the number of frames (F) and the sampling period (T). During this experiment, the UniFormerS was pre-trained on ImageNet, and the 3D U-Net was trained from scratch.

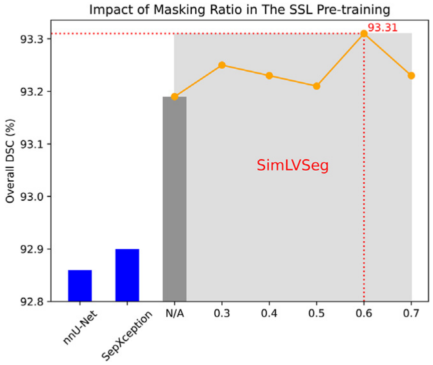


Figure 8. Impact of masking ratio to SimLVSeg-SI. The optimum masking ratio is 60%. N/A, without pre-training.

the two 95% CIs, indicating statistical significance. Moreover, the enhanced segmentation performance observed in the middle frame indicated that the SSL pre-trained model was particularly good at leveraging temporal dynamics, suggesting the model effectively utilized the contextual information from the frames preceding and following the middle frame to improve its segmentation accuracy. Furthermore, while the ES DSC with SSL was higher, the ED DSC was lower than that of the network without SSL pre-training. In CAMUS, the ED and ES frames were located at the sequence edges, limiting temporal context from their preceding or following frames.

Discussion

Table 1 shows that while being more efficient, SimLVSeg outperformed the highest reported DSC on the EchoNet-Dynamic test set. SimLVSeg video networks aggregated both spatial and temporal information by analyzing multiple echocardiogram frames in a single pass. The networks predicted an LV segmentation trace for every input frame at once, thus eliminating redundancy in analyzing the same frames multiple times as in Thomas et al. [20] and Wu et al. [24]. In addition, the SimLVSeg training pipeline was simple yet effective, easy to implement and scalable, as it did not require pseudo labels [25,46] or temporal regularization [28]. Compared with [25,46], SimLVSeg did not depend on a specific heart stage, thus eliminating the burden of locating the ED and ES frame when creating training data. This also allowed us to easily leverage non-ED and -ES frames for supervision if their corresponding segmentation labels were available. Figure 7 highlights the robustness of SimLVSeg to the sampling hyper-parameters. This allowed for a broader design space to meet hardware limitations such as memory and computing power (floating point operations per second) while still achieving a satisfactory segmentation performance.

We observed that randomly masking a significant portion (60%) of an echocardiogram clip during SSL pre-training resulted in the best performance. Masking the SSL improved the overall DSC of the SI approach from 93.19% to 93.31%, as reported in Figure 8. Further, as shown in Figure 9, we observed that self-supervision with temporal masking enabled the network to maintain better temporal consistency across predictions in a given echocardiogram clip. Figure 9a demonstrates that a video segmentation model pre-trained with self-supervised temporal masking is more resilient to noise and missing artifacts. The pre-training stage also alleviates the over-segmentation and temporal inconsistency issues that are commonly encountered in echocardiography caused by unclear (or, even worse, invisible) boundaries. Additionally, the SSL pre-trained model achieved a smoother LV segmentation area prediction with significantly less rapid fluctuations (Fig. 9b), indicating better temporal consistency. We further investigated this phenomenon in the frequency domain by applying fast Fourier transform to the predicted signals (LV area), as depicted in Figure 9c. We observed that SSL pre-training resulted in a lower magnitude of the high-frequency components, which are typically the result of noise and rapid fluctuation. Based on these observations, we hypothesize that the pre-training stage helps the 3-D U-Net model to better learn the semantic features that are useful for estimating human heart structures in the apical four-chambers view, resulting in a more robust prediction. Additionally, Table 3 showcases the significant impact of the SSL pre-training stage when testing with samples subject to distribution shifts, indicating that SSL pre-training enhances the model's generalization capability. These findings indicate that pre-training with self-supervision remarkably benefits the downstream LV segmentation task. Hence, SSL with vast echocardiogram videos can be a promising solution to provide strong pre-trained models that can generalize well in downstream echocardiography-related clinical tasks.

We have shown that both the SI and 3-D segmentation networks trained using our proposed SimLVSeg are capable of accurately segmenting the LV in echocardiogram videos. Both SimLVSeg-3-D and SimLVSeg-SI outperform the state of the art [18], suggesting that the superior performance could be attributed to the SimLVSeg design rather than selection of the underlying network architectures. 3-D U-Net performance was slightly better than the SI network with the UniFormer-S backbone. However, designing a backbone for 3-D U-Net is not straightforward as it requires tedious hyperparameter tuning. On the other hand, there are plenty of optimized models that can be utilized as a backbone for the SI approach. For instance, MobileNetV3, with only 6.69 M of parameters, can give an on-par performance with 93.16% overall DSC, as seen in Table 2. The pre-trained models on ImageNet can also help generalize better if we only have a small amount of data. Moreover, many SSL algorithms for 2-D can also be explored to further improve SimLVSeg performance.

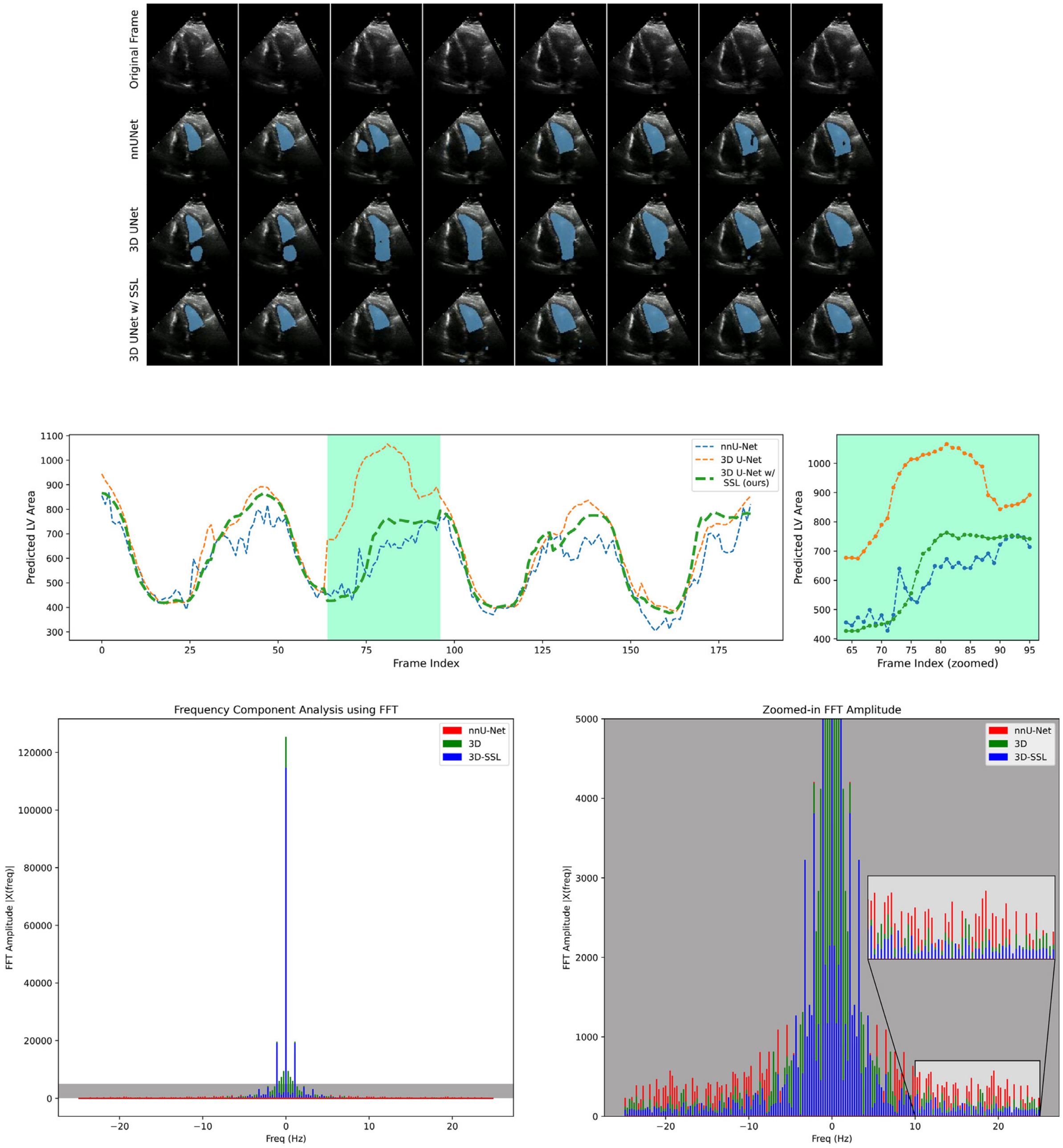


Figure 9. Qualitative results for the performance of nnU-Net (2-D) [14], 3-D U-Net and 3-D U-Net with self-supervision (SimLVSeg-3D) without any post-processing trick against a challenging case where the Mitral valve is unclear. (a) We observe that the SimLVSeg-3D is more resilient to noise and missing artifacts in the input frames. (b) The predicted LV segmentation area is smoother and more consistent for the SimLVSeg-3D compared with others. A zoomed version of the plot is shown on the right. (c) Frequency analysis using fast Fourier transform on the predicted left ventricular areas shows a lower magnitude of high-frequency components (*i.e.*, lower noise) in the SimLVSeg-3D compared with others.

Table 2
Ablation study on various encoder backbones

Approach (# frames, period)	Backbone	% DSC (overall)	Parameters (M)	FLOPs (G)	
				Single pass	One frame
SI (16, 5)	MobileNetV3	93.16	6.69	12.46	0.78
	ResNet-18	93.23	14.33	21.75	1.36
	ViT-B/16	92.98	89.10	120.20	7.51
3-D (32, 1)	3-D U-Net-S	93.27	11.26	27.34	0.85

The approach was robust for the selection of backbone complexity. The SI backbones were pre-trained on the ImageNet dataset, while the 3D U-Net-S was trained from scratch.
DSC, dice similarity coefficient; G, Gigaflops; M, Millions of parameters; SI, super image.

Table 3
SimLVSeg-3D performance on the CAMUS dataset (out of distribution)

SSL	DSC (95% CI)			
	Overall	Middle	ES	ED
X	0.9044 (0.9039−0.9050)	0.8976 (0.8952−0.8999)	0.8901 (0.8875−0.8926)	0.9234 (0.9217−0.9251)
✓	0.9062 (0.9057−0.9067)	0.9013 (0.8990−0.9034)	0.8949 (0.8924−0.8973)	0.9155 (0.9138−0.9172)

Using the proposed self-supervised temporal masking for pre-training leads to better generalization.
DSC, dice similarity coefficient; ED, end diastole; ES, end systole; SSL, self-supervised learning.

Conclusion and future work

Here we proposed a novel paradigm to tackle the LV segmentation task on echocardiogram videos, namely SimLVSeg, and our method outperformed other works on the EchoNet-Dynamic test set. SimLVSeg utilizes a video segmentation network that efficiently combines both spatial and temporal information. The network is pre-trained on a reconstruction task and then fine-tuned with sparse annotations to predict LV. An extensive experiment was performed to show the superiority of SimLVSeg, both quantitatively and qualitatively. We expect that this work will motivate researchers to explore more of the video segmentation approach for LV instead of working on frame-by-frame prediction.

Despite SimLVSeg's remarkable performance for consistent LV segmentation, we limited our experiments to self-supervision using temporal masking only. However, there remains scope to improve self-supervision pre-training by identifying the optimum masking scheme between existing masking strategies [34,47], such as temporal, random spatiotemporal, space-wise and block-wise masking. Additionally, this work only considered LV segmentation from a single echocardiography view, *i.e.*, the apical four-chamber view. Extending SimLVSeg for LV segmentation from multi-view echocardiogram videos can further improve the overall performance and its usage in clinical practice.

Conflict of interest

The authors declare no competing interests.

Data availability statement

The datasets used in this work are publicly available. The preprocessed datasets may be obtained on request to the authors.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT in order to enhance writing and improve language. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

[1] Horgan SJ, Uretsky S. Essential Echocardiography: A Companion to Braunwald's Heart Disease Echocardiography in the context of other cardiac imaging modalities. Amsterdam, Netherlands: Elsevier; 2019. p. 460–73.
[2] Ford MK, Beattie WS, Wijeysundera DN. Systematic review: prediction of perioperative cardiac complications and mortality by the revised cardiac risk index. Ann Intern Med 2010;152(1):26–35.
[3] Folland ED, Parisi AF, Moynihan PF, Jones DR, Feldman CL, Tow DE. Assessment of left ventricular ejection fraction and volumes by real-time, two-dimensional echocardiography. A comparison of cineangiographic and radionuclide techniques. Circulation 1979;60(4):760–6.
[4] Heidenreich PA, Trogdon JG, Khavjou OA, Butler J, Dracup K, Ezekowitz MD, et al. Forecasting the future of cardiovascular disease in the United States: a policy statement from the American Heart Association. Circulation 2011;123(8):933–44.
[5] Lang RM, Badano LP, Mor-Avi V, Afilalo J, Armstrong A, Ernande L, et al. Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging. Eur Heart J Cardiovasc Imaging 2015;16(3):233–70.
[6] Mada RO, Lysyansky P, Daraban AM, Duchenne J, Voigt JU. How to define end-diastole and end-systole?: Impact of timing on strain measurements. JACC Cardiovasc Imaging 2015;8(2):148–57.
[7] Ouyang D, He B, Ghorbani A, Yuan N, Ebinger J, Langlotz CP, et al. Video-based AI for beat-to-beat assessment of cardiac function. Nature 2020;580(7802):252–6.
[8] Farsalinos KE, Daraban AM, Unlu S, Thomas JD, Badano LP, Voigt JU. Head-to-head comparison of global longitudinal strain measurements among nine different vendors: The EACVI/ASE inter-vendor comparison study. J Am Soc Echocardiogr 2015;28(10) 1171−81.e2.
[9] Cole GD, Dhutia NM, Shun-Shin MJ, Willson K, Harrison J, Raphael CE, et al. Defining the real-world reproducibility of visual grading of left ventricular function and visual estimation of left ventricular ejection fraction: impact of image quality, experience and accreditation. Int J Cardiovasc Imaging 2015;31(7):1303–14.
[10] Pellikka PA, She L, Holly TA, Lin G, Varadarajan P, Pai RG, et al. Variability in ejection fraction measured by echocardiography, gated single-photon emission computed tomography, and cardiac magnetic resonance in patients with coronary artery disease and left ventricular dysfunction. JAMA Netw Open 2018;1(4):e181456.
[11] Kang S, Kim SJ, Ahn HG, Cha K-C, Yang S. Left ventricle segmentation in transesophageal echocardiography images using a deep neural network. PLoS One 2023;18 (1):e0280485.
[12] Ghorbani A, Ouyang D, Abid A, He B, Chen JH, Harrington RA, et al. Deep learning interpretation of echocardiograms. NPJ Digit Med 2020;3(1):10.
[13] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, editors. MICCAI 2015. Cham, Switzerland: Springer International Publishing; 2015. p. 234–41.
[14] Isensee F, Jaeger PF, Kohl SA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods 2021;18(2):203–11.
[15] Smistad E, Ostvik A, Haugen BO, Lpvstakken L. 2D left ventricle segmentation using deep learning. In: Paper presented at: IEEE International Ultrasonics Symposium (IUS); September 6−9, 2017. p. Washington, DC, USA1–4.
[16] Hu Y, Guo L, Lei B, Mao M, Jin Z, Elazab A, et al. Fully automatic pediatric echocardiography segmentation using deep convolutional networks based on bisenet. In: Paper presented at: 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); 23−27 July, 2019. p. Berlin, Germany6561–4.
[17] Leclerc S, Smistad E, Grenier T, Lartizien C, Ostvik A, Cervenansky F, et al. RU-Net: A refining segmentation network for 2D echocardiography. In: Paper presented at: IEEE International Ultrasonics Symposium (IUS); October 6−9, 2019. p. Glasgow, Scotland1160–3.
[18] Chen E, Cai Z, Lai J-H, et al. Weakly supervised semantic segmentation of echocardiography videos via multi-level features selection. In: Yu S, Zhang Z, Yuen PC, Han J, Tan T, Guo Y, editors. Pattern Recognition and Computer Vision. Switzerland: Springer Nature; 2022. p. 388–400.
[19] Reddy CD, Lopez L, Ouyang D, Zou JY, He B. Video-based deep learning for automated assessment of left ventricular ejection fraction in pediatric patients. J Am Soc Echocardiogr 2023;36(5):482–9.
[20] Thomas S, Gilbert A, Ben-Yosef G. Light-weight spatio-temporal graphs for segmentation and ejection fraction prediction in cardiac ultrasound. In: Wang L, Dou Q, Fletcher PT, Speidel S, Li S, editors. MICCAI 2022. Switzerland: Springer Nature; 2022. p. 380–90.
[21] Li M, Zhang W, Yang G, Wang C, Zhang H, Liu H, et al. Recurrent aggregation learning for multi-view echocardiographic sequences segmentation. Paper presented at: MICCAI 2019: 22nd International Conference, Part II 22. Shenzhen, China: Springer; October 13−17, 2019. p. 678–86.
[22] Ahn SS, Ta K, Thorn S, Langdon J, Sinusas AJ, Duncan JS. Multi-frame attention network for left ventricle segmentation in 3D echocardiography. Paper presented at: MICCAI 2021: Part I 24. Strasbourg, France: Springer; September 27−October 1, 2021. p. 348–57.
[23] Sirjani N, Moradi S, Oghli M, et al. Automatic cardiac evaluations using a deep video object segmentation network. Insights Imaging 2022;13:69.
[24] Wu H, Liu J, Xiao F, Wen Z, Cheng L, Qin J. Semi-supervised segmentation of echocardiography videos via noise-resilient spatiotemporal semantic calibration and fusion. Med Image Anal 2022;78:102397.
[25] Wei H, Ma J, Zhou Y, Xue W, Ni D. Co-learning of appearance and shape for precise ejection fraction estimation from echocardiographic sequences. Med Image Anal 2023;84:102686.
[26] Chen Y, Zhang X, Haggerty CM, Stough JV. Assessing the generalizability of temporally coherent echocardiography video segmentation. Medical Imaging 2021: Image Processing, Vol. 11596. Bellingham, WA, USA: SPIE; 2021. p. 463–9.

[27] Stough JV, Raghunath S, Zhang X, Pfeifer JM, Fornwalt BK, Haggerty CM. Left ventricular and atrial segmentation of 2D echocardiography with convolutional neural networks. Medical Imaging: Image Processing. Bellingham, WA, USA: SPIE; 2020.
[28] Painchaud N, Duchateau N, Bernard O, Jodoin P-M. Echocardiography segmentation with enforced temporal consistency. IEEE Trans Med Imaging 2022;41(10):2867–78.
[29] Yang J, Ding X, Zheng Z, Xu X, Li X. Graphecho: Graph-driven unsupervised domain adaptation for echocardiogram video segmentation. In: Paper presented at: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France. IEEE Computer Society; October 2−3, 2023. p. 11844–53.
[30] Dezaki FT, Luong C, Ginsberg T, Rohling R, Gin K, Abolmae-sumi P, et al. Echo-syncnet: Self-supervised cardiac view synchronization in echocardiography. IEEE Trans Med Imaging 2021.
[31] Dai W, Li X, Ding X, Cheng K-T. Cyclical self-supervision for semi-supervised ejection fraction prediction from echocardiogram videos. IEEE Trans Med Imaging 2023;42 (5):1446–61.
[32] Saeed M, Muhtaseb R, Yaqub M. Contrastive pretraining for echocardiography segmentation with limited data. In: Paper presented at: Medical Image Understanding and Analysis: 26th Annual Conference, MIUA 2022, Proceedings. Cambridge, UK. Springer; July 27−29, 2022. p. 680–91.
[33] Feichtenhofer C, Fan H, Li Y, He K. Masked autoencoders as spatiotemporal learners. In: Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K, Oh A, editors. Advances in Neural Information Processing Systems, Vol. 35. New York: Curran Associates, Inc.; 2022. p. 35946–58.
[34] Tong Z, Song Y, Wang J, Wang L. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. arXiv 2022 2203.12602.
[35] Cicek O, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In: Paper presented at: MICCAI 2016: Part II 19. Athens, Greece. Springer; October 17−21, 2016. p. 424–32.
[36] Fan Q, Chen C-F, Panda R. Can an image classifier suffice for action recognition? Paper presented at: International Conference on Learning Representations; April 25 −29, 2022. virtual.
[37] Sobirov I, Saeed N, Yaqub M. Super images − a new 2D perspective on 3D medical imaging analysis. In: Waiter G, Lambrou T, Leontidis G, Oren N, Morris T, Gordon S, editors. Medical Image Understanding and Analysis. Cham, Switzerland: Springer Nature; 2024. p. 325–37.
[38] Kerfoot E, Clough J, Oksuz I, Lee J, King AP, Schnabel JA, et al. Left-ventricle quantification using residual U-Net. In: Pop M, Sermesant M, Zhao J, Li S, McLeod K, Young A, editors. Statistical Atlases and Computational Models of the Heart. Atrial Segmentation and LV Quantification Challenges. Cham, Switzerland: Springer International Publishing; 2019. p. 371–80.
[39] Li K, Wang Y, Gao P, Song G, Liu Y, Li H, et al. Uniformer: Unified transformer for efficient spatial-temporal representation learning. ICLR; 2022.
[40] Muhtaseb R, Yaqub M. Echocotr: Estimation of the left ventricular ejection fraction from spatiotemporal echocardiography. In: Wang L, Dou Q, Fletcher PT, Speidel S, Li S, editors. MICCAI 2022. Cham, Switzerland: Springer Nature; 2022. p. 370–9.
[41] Leclerc S, Smistad E, Pedrosa J, Stvik A, Cervenansky F, Espinosa F, et al. Deep learning for segmentation using an open large-scale dataset in 2D echocardiography. IEEE Trans Med Imaging 2019;38(9) 219−210.
[42] Isensee F, Wald T, Ulrich C, Baumgartner M, Roy S, Maier-Hein K, et al. nnU-Net revisited: A call for rigorous validation in 3D medical image segmentation. arXiv 2024 2404.09556.
[43] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Paper presented at: 2016 IEEE Conference on CVPR; June 27−30, 2016. p. Las Vegas, NV770–8https://api.semanticscholar.org/CorpusID:206594692.
[44] Howard A, Pang R, Adam H, Le QV, Sandler M, Chen B, et al. Searching for MobileNetV3. In: Paper presented at: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019; October 27−November 2, 2019. p. Seoul, Korea1314–24.
[45] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: transformers for image recognition at scale. In: Paper presented at: 9th International Conference on Learning Representations, ICLR 2021; May 3−7, 2021.
[46] Wei H, Cao H, Cao Y, Zhou Y, Xue W, Ni D, et al. Temporal-consistent segmentation of echocardiography with co-learning from appearance and shape. In: Martel AL, Abolmaesumi P, Stoyanov D, Mateus D, Zuluaga MA, Zhou SK, editors. MICCAI 2020. Cham, Switzerland: Springer International Publishing; 2020. p. 623–32.
[47] Wang L, Huang B, Zhao Z, Tong Z, He Y, Wang Y, et al. VideoMAE V2: scaling video masked autoencoders with dual masking. In: Paper presented at: IEEE/CVF Conference on CVPR; June 18−22, 2023. p. Vancouver, BC14549–60.