Semi-supervised Echocardiography Video Segmentation via Anchor Semantic Awareness and Continuous Pseudo-label Reforging

Yunpeng Fang[1], Yimu Sun[1], Jingxing Guo[1], Huisi Wu[1]*, Jing Qin[2]
[1]College of Computer Science and Software Engineering, Shenzhen University
[2]Centre for Smart Health, School of Nursing, The Hong Kong Polytechnic University
2410103037@mails.szu.edu.cn, hswu@szu.edu.cn

Abstract

Automatic and accurate echocardiography video segmentation is essential for efficient and repeatable measurements of key clinical functional indicators for the diagnosis of cardiovascular diseases. However, it is an extremely challenging task to obtain high-quality segmentation results throughout the cardiac cycle owing to (1) the inherent speckle noise in echocardiography videos, (2) the complex dynamic motions of cardiac structures, and (3) the scarcity of annotated data. To comprehensively address these challenges, we propose a novel semi-supervised model, EchoForge, which can achieve accurate and real-time echocardiography video segmentation with very limited annotations. EchoForge introduces an Anchor Semantic Awareness (ASA) module that refines ambiguous regions using learnable anchors and propagates structural prototypes across frames to enhance boundary delineation and temporal consistency. Building upon ASA, a Continuous Pseudo-label Reforging (CPR) module progressively integrates and refines pseudo-labels via channel-wise attention, providing robust supervision. Extensive experiments on the CAMUS and EchoNet-Dynamic benchmarks demonstrate that EchoForge outperforms state-of-the-art (SOTA) methods in accuracy while maintaining real-time efficiency. The code is available at https://github.com/YunPeng-Fang/EchoForge.

1. Introduction

Echocardiography serves as one of the first-line tools for cardiac assessment in clinical practice owing to its merits of real-time imaging, non-invasiveness, and cost-effectiveness [1]. Accurate segmentation of key anatomical structures, such as the left ventricular (LV) endocardium, from echocardiography videos is critical for quantifying cardiac function metrics, including ejection fraction (EF), end-diastolic volume (EDV), and end-systolic volume (ESV). These metrics

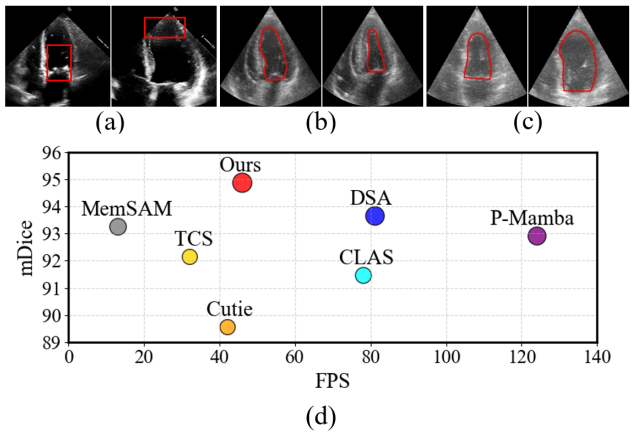


Figure 1. The challenges of echocardiography video segmentation and performance comparison: (a) speckle noise and blurred boundaries; (b)-(c) variations of the left ventricle across frames; (d) accuracy vs. efficiency comparison with SOTA methods on the CAMUS test set.

are essential for cardiovascular disease (CVD) diagnosis and treatment planning [5]. However, manual segmentation by clinicians is labor-intensive, time-consuming, and subject to inter-observer variability. In this regard, there is an urgent need for automated solutions to improve both accuracy and efficiency [24, 28].

Despite its clinical significance, automatic echocardiography video segmentation is hindered by several inherent difficulties. First, as illustrated in Figure 1 (a), ultrasound images are intrinsically plagued by speckle noise and artifacts, which obscure anatomical details and render the boundaries of target structures ambiguous [4, 34]. Second, the shape and scale of cardiac structures show significant spatiotemporal variations during contraction and relaxation (Figure 1 (b-c)). Third, the labor-intensive nature of manual delineation makes ground truth annotations typically sparse, often restricted to only the end-diastolic (ED) and end-systolic (ES) frames. This scarcity of supervision severely constrains

*Corresponding Author

model training. Finally, the need for real-time segmentation and measurement in clinical practice also poses a challenge to computational efficiency.

In recent years, many deep learning approaches have been proposed to address these challenges. Early investigations focus on 2D convolutional neural networks trained on the sparsely annotated keyframes [33, 35]. However, these methods largely ignore the temporal consistency of cardiac motion and hence are susceptible to speckle noise and shape variations. To improve segmentation accuracy, some other studies take cardiac motion into account and harness optical flow to enhance temporal coherence [18, 22]. However, these methods are still quite sensitive to speckle noise and artifacts, leading to unsatisfactory segmentation results. More recently, the advent of large foundation models, particularly SAM [21], has opened a new avenue, offering powerful representational capabilities to deal with noise, artifacts, and the scarcity of labeled data [9, 19, 27, 45]. Nevertheless, directly using SAM for video segmentation cannot achieve satisfactory performance, as it is incapable of capturing crucial temporal dynamics.

In this paper, we propose a novel semi-supervised echocardiography video segmentation model, named EchoForge, to comprehensively address the aforementioned challenges. Our model is composed of two major innovative modules: Anchor Semantic Awareness (ASA) and Continuous Pseudo-label Reforging (CPR). The ASA further includes two components: Anchor Recalibration (ARC) and Temporal Semantic Fusion (TSF). Specifically, we first propose an ARC scheme, which is centered on learnable anchors and flexibly adjusts the features of the model's uncertain areas, thereby effectively suppressing speckle noise interference and enhancing boundary structure consistency. Based on ARC, we propose the TSF, which extracts semantic prototypes using contextual references and injects them into the current frame via the attention mechanism to further improve boundary precision and maintain spatiotemporal consistency. Furthermore, based on ASA, the CPR module progressively integrates high-quality pseudo-labels through lightweight channel-wise attention, and continuously reforges them to provide more robust supervision. We conducted extensive experiments on two benchmarking datasets, CAMUS [23] and EchoNet-Dynamic [31], to validate the effectiveness of the proposed model. As shown in Figure 1 (d), our method achieves better performance than SOTA methods while maintaining real-time efficiency. Our major contributions can be summarized as follows:

- We propose EchoForge, a novel echocardiography video segmentation framework, which includes two innovative modules: an Anchor Semantic Awareness (ASA) module and a Continuous Pseudo-label Reforging (CPR) module.
- The ASA module can flexibly adjust the features of the model's uncertain regions and propagate semantic prototypes, thereby suppressing speckle noise and enhancing boundary consistency. Building upon ASA, the CPR module progressively integrates and reforges high-quality pseudo-labels, thus achieving robust supervision.
- EchoForge achieves SOTA segmentation performance on two well-known benchmarks while maintaining real-time inference speeds.

2. Related Work

2.1. Echocardiography Video Segmentation

Research in echocardiography video segmentation has primarily evolved along several distinct lines to address its challenges [32]. Foundational approaches often employ 2D Convolutional Neural Networks (CNNs) for frame-by-frame analysis [2, 6, 14, 28, 43]. While straightforward, such methods inherently disregard the temporal dimension, often resulting in predictions that lack inter-frame consistency. To address this, a subsequent line of work integrated optical flow to enforce temporal coherence [13, 17, 30]. However, optical flow is notoriously sensitive to the low signal-to-noise ratio and heavy speckle noise characteristic of ultrasound, frequently leading to inaccurate pixel-level motion fields. Recognizing these limitations, another prominent strategy involves fusing features from adjacent frames [16, 29, 38, 40, 44]. While this improves short-range consistency, these methods typically operate within a limited temporal window and struggle to capture the full, long-range dynamics of the cardiac cycle. In addition, recent methods [36, 39] also use pseudo-label to assist sparsely labeled echocardiography segmentation, however, these methods suffer from a critical flaw where initial errors are propagated and accumulated, degrading the overall model performance.

2.2. Semi-supervised Video Segmentation

Semi-supervised video segmentation has emerged as a dominant paradigm for segmenting objects in video, tasked with propagating an initial reference mask throughout the sequence [11]. Simply fusing various temporal features may compromise their completeness and validity of the information, resulting in degraded segmentation performance. To address this issue, on the one hand, memory-based networks are proposed, which maintain and query a memory bank of past object features to guide the segmentation of subsequent frames [3, 7, 8]. However, the scarcity of data in echocardiography video datasets makes it difficult for the aforementioned semi-supervised networks to be directly applied to the task of echocardiography video segmentation. On the other hand, teacher-student models and cross-pseudo-supervision (CPS) are also at the core of these efforts [37, 41]. The former generates pseudo-labels from a momentum-updated teacher network, while the latter enforces consistency between two parallel student models. However, the separate training flows

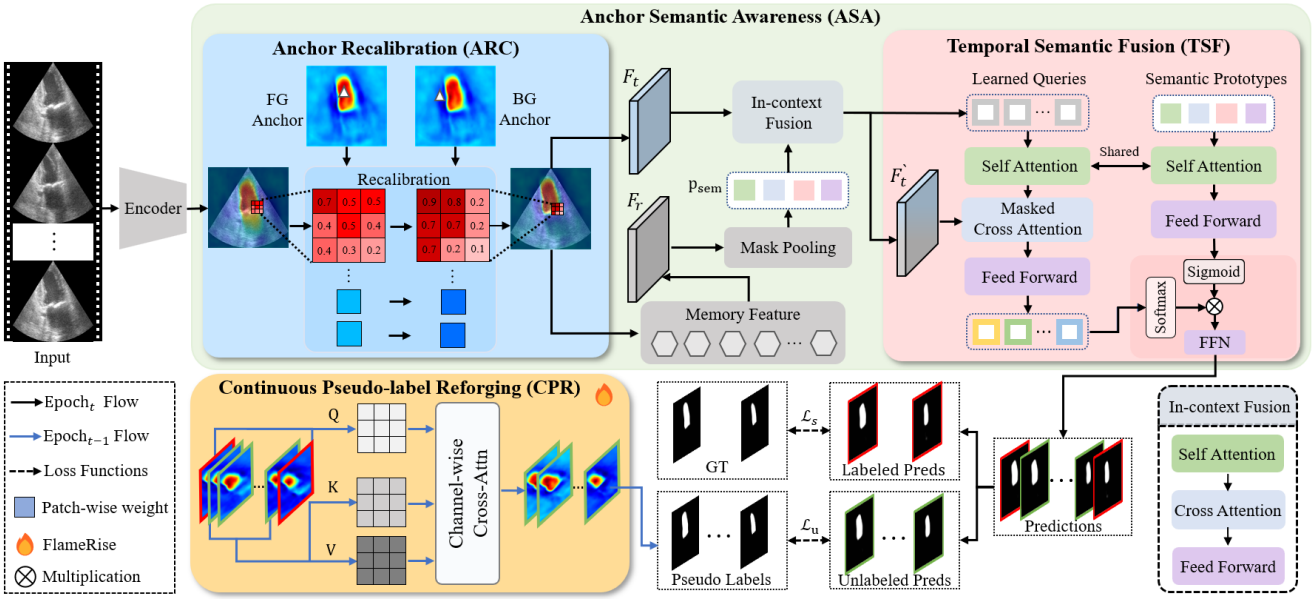


Figure 2. Overview of the proposed EchoForge framework, consisting of the ASA and CPR modules. The ASA module utilizes learnable anchors and semantic awareness to suppress speckle noise, while the CPR module leverages annotated frames to reforge pseudo-labels, providing robust supervision.

in these frameworks cause the model to be heavily biased by the annotated keyframes, thereby failing to effectively leverage the unlabeled frames to learn robust representations and temporal variations.

3. Methods

3.1. Overview

The overall architecture of our proposed EchoForge framework is illustrated in Figure 2. It is a novel semi-supervised echocardiography video segmentation model comprising two main components: Anchor Semantic Awareness (ASA) and Continuous Pseudo-label Reforging (CPR). To mitigate the impact of speckle noise in ultrasound images, we first design an Anchor Recalibration (ARC) module within ASA to recalibrate features via learnable anchors, helping the model refine ambiguous regions. Furthermore, given that cardiac structures exhibit significant shape variations across frames in echocardiography videos, we introduce a Temporal Semantic Fusion (TSF) module. This module improves the spatiotemporal consistency of the segmentation by propagating semantic prototypes. Building upon ASA, to fully exploit the information from unlabeled frames, we propose the CPR module. It leverages features from labeled frames to continuously reforge pseudo-labels for the unlabeled ones, progressively improving their quality. We detail each component in the subsequent sections.

3.2. Anchor Semantic Awareness

3.2.1. Anchor Recalibration

Ultrasound images are intrinsically plagued by speckle noise; thus, traditional global attention mechanisms are easily distracted by the noise, resulting in inaccurate segmentation masks. To combat this, we propose the ARC module. Unlike previous object detection works, the anchors here do not refer to general candidate boxes, but rather to a set of learnable feature vectors that carry preliminary foreground and background information. They act as magnets to attract the feature blocks most similar to the target within the complex and noisy ultrasound background.

During anchor initialization, we first obtain the weight distribution of the foreground and background anchors by applying a pixel-wise 1×1 convolution mapping and channel-level softmax to the encoder output feature map $X \in \mathbb{R}^{C\times H\times W}$:

$$M_i(x,y) = \frac{\exp\big(f_i(X)_{x,y}\big)}{\sum_{j=1}^{2}\exp\big(f_j(X)_{x,y}\big)}, \quad i \in \{1,2\} \tag{1}$$

where f_i represents the similarity score map corresponding to the i-th anchor. Subsequently, we use global weighted average pooling to aggregate X and extract the initial anchor representation $a_i^{(0)} \in \mathbb{R}^C$:

$$a_i^{(0)} = \frac{\sum_{x,y} M_i(x,y)\, X(x,y)}{\sum_{x,y} M_i(x,y)}. \tag{2}$$

The anchor update process is illustrated in Figure 3. After obtaining the initial anchor $a_i^{(0)}$, we introduce the KNN

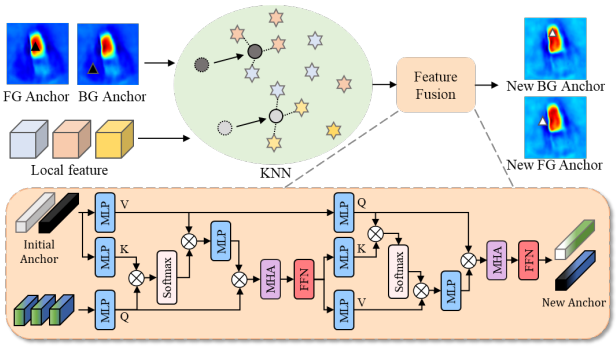


Figure 3. Detailed illustration of the anchor update process. Initial anchors are first filtered with local feature via KNN, and then updated anchors are generated through a Feature Fusion module.

algorithm for neighbor feature screening. Specifically, the encoder feature map X is first flattened into a set of pixel features, and then for each initial anchor $a_i^{(0)}$, the K pixels with the smallest distance to the anchor in the feature space are selected to form the neighbor set N_i. Next, the set N_i is input into the Feature Fusion (FF) module together with the current anchor representation $a_i^{(0)}$. The FF module uses N_i and $a_i^{(0)}$ as queries to perform cross-representation attention, ultimately generating the new anchor a_i through a feed-forward network and residual connection.

Finally, the anchor recalibration process is detailed in Figure 2. The feature map X is divided into several non-overlapping patches P_k. We compute their cosine similarities with both anchors to obtain the foreground probability s_k^{FG} and the background probability s_k^{BG}. For patches with high confidence, the original features are retained. For patches falling within the uncertainty interval [0.4,0.6], their weights are dynamically determined based on the similarity difference, and the patch features are linearly interpolated towards the anchor a_i with the higher confidence. After this process, the updated feature map F_t is obtained.

3.2.2. Temporal Semantic Fusion

In order to handle the significant shape variations of the left ventricle across frames in echocardiography videos, we propose the Temporal Semantic Fusion (TSF) module based on ARC. The core of TSF is to extract and propagate key anatomical prototypes to improve the temporal consistency of the model. As shown in Figure 2, for the reference feature map $F_r \in \mathbb{R}^{C\times H\times W}$ extracted from the memory feature and its associated mask annotation $\{m_i^r\}_{i=1}^N$, we first perform mask pooling to extract a set of semantic tags $t_{\text{sem},i} \in \mathbb{R}^C$:

$$t_{\text{sem},i} = \frac{1}{\sum_{u,v} m_i^r(u,v)} \sum_{u,v} m_i^r(u,v)\, F_r(u,v), \quad (3)$$

which are collected into $T_{\text{sem}} = [t_{\text{sem},1}; \ldots; t_{\text{sem},N}] \in \mathbb{R}^{N\times C}$. Simultaneously, the target frame $F_t \in \mathbb{R}^{C\times H\times W}$ is flattened into patch tokens $X_t \in \mathbb{R}^{HW\times C}$. We introduce an In-context Fusion module to model the correlation between the reference and target image features. The semantic tokens T_{sem} serve as queries while X_t provides the keys and values. The process of this module can be summarized as follows:

$$\langle P_{sem},\ F_t' \rangle = \text{In-context Fusion}\,(T_{sem},\ F_t;\ \theta)\,, \quad (4)$$

where θ refers to the parameters of the In-context Fusion module. This module is constructed as a Transformer block, comprising a self-attention mechanism, cross-attention, and a feed-forward network. The module is responsible for fusing the tokens T_{sem} with the target feature F_t. Specifically, within the cross-attention process, these tokens and the target feature serve as keys and values for one another. Through this fusion, the enhanced target feature $F_t' \in \mathbb{R}^{C\times H\times W}$ and semantic prototypes $P_{sem} \in \mathbb{R}^{N_q\times C}$ can be obtained.

Next, we introduce attention structures for deep semantic interaction between learned queries $Q \in \mathbb{R}^{N_q\times C}$ and semantic prototypes P_{sem}. We first perform self-attention updates $Q' = \text{SelfAttn}(Q)$ and $\hat{P} = \text{SelfAttn}(P_{sem})$. Using the fused feature F_t' as the value, the query branch performs masked cross-attention $\hat{Q} = \text{CrossAttn}(Q')$. Finally, the representations enter a standard feed-forward network (FFN) to yield $Q_{final} = \text{FFN}(\hat{Q})$ and $P_{final} = \text{FFN}(\hat{P})$, which are combined to generate the prediction masks.

3.3. Continuous Pseudo-label Reforging

Labeled data are scarce in echocardiography videos; therefore, to fully utilize the information of intermediate frames, we introduce a pseudo-labeling strategy. However, the pseudo-labels generated by existing methods are usually of poor quality, especially in the early stages of training. To alleviate this problem, based on ASA, we introduce Continuous Pseudo-label Reforging (CPR), a lightweight attention module designed to refine noisy pseudo-labels by semantically reforging them from labeled features.

The structure of CPR is shown in Figure 2. The prediction features generated by the model can be divided into labeled frame features $F^L \in \mathbb{R}^{C\times H\times W}$ and unlabeled frame features $F^U \in \mathbb{R}^{C\times H\times W}$. Then, we use F^L as the query, F^U as the key and value to calculate the channel-level cross-attention. The specific process can be described as follows:

$$A = \text{softmax}(\text{IN}(Q^T K)), \quad (5)$$

$$\hat{F}^U = AV^T, \quad (6)$$

where $Q = \text{flatten}(F^L)W_q$, $K = \text{flatten}(F^U)W_k$, and $V = \text{flatten}(F^U)W_v$. Finally, the reconstructed feature $\hat{F}^U$ is passed through a semantic alignment to obtain the new pseudo-labels $\hat{y}^U$.

FlameRise. Despite our proposed CPR, training with pseudo labels at all epochs still causes the model to overfit

to early noisy predictions. To address this issue, we propose FlameRise, a training strategy that gradually increases the contribution of pseudo-label supervision as training progresses, like a flame rising. Concretely, let E denote the total number of epochs and $e \in [1, E]$ the current epoch index. We define two time-dependent schedules. First, the pseudo-label weight $\lambda(e)$, which controls the relative importance of the unsupervised loss:

$$\lambda(e) = \begin{cases} 0, & e \le E_0, \\ \lambda_{\max} \dfrac{e - E_0}{E_1 - E_0}, & E_0 < e \le E_1, \\ \lambda_{\max}, & e > E_1, \end{cases} \quad (7)$$

where E_0 is the burn-in epoch before any pseudo-labels are used, E_1 is the ramp-up endpoint, and $\lambda_{\max}$ is the maximum unsupervised weight. Second, the confidence threshold $\tau(e)$, which determines which pixels in an unlabeled frame are considered trustworthy:

$$\tau(e) = \begin{cases} \tau_0, & e \le E_0, \\ \tau_0 - (\tau_0 - \tau_1) \dfrac{e - E_0}{E_1 - E_0}, & E_0 < e \le E_1, \\ \tau_1, & e > E_1, \end{cases} \quad (8)$$

where τ_0 is the initial threshold and τ_1 the final threshold. At each iteration, for a batch of unlabeled frames, we predict softmax scores at each pixel location and construct a mask. The unsupervised loss is then computed only over high-confidence locations.

3.4. Loss Function

We use the Dice loss $\mathcal{L}_{dice}$ to supervise our model training. In addition, as mentioned in the previous training strategy, we use loss $\mathcal{L}_{U(e)}$ to supervise the unlabeled frames. To further refine boundary details, we introduce the binary cross-entropy loss $\mathcal{L}_{bce}$. Therefore, our total loss function can be defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{bce}}(P_i, G_i) + \mathcal{L}_{\text{dice}}(P_i, G_i) + \mathcal{L}_{U(e)}(P_i, \hat{y}^U), \quad (9)$$

where P_i denotes the predictions for the frames, G_i represents the corresponding ground truth and $\hat{y}^U$ stands for the pseudo-labels.

4. Experiments

4.1. Datasets and Evaluation Metrics

We evaluated our method on two public echocardiography video datasets: CAMUS [23] and EchoNet-Dynamic[31].
CAMUS consists of 500 patient cases, each containing apical two-chamber and four-chamber echocardiography videos. CAMUS provides annotations for all video frames.

EchoNet-Dynamic consists of 10,030 apical four-chamber echocardiography videos, each cropped to exclude any information beyond the ultrasound scan sector. EchoNet-Dynamic only provides annotations for the ED and ES frames.

Due to the scarcity of annotations in other datasets, we only use the annotations of ED and ES frames provided in CAMUS for semi-supervised training. We derived two variants from the CAMUS dataset: CAMUS-Semi and CAMUS-Full. During training, models on both variants are trained using only the annotations of the ED and ES frames. During testing, CAMUS-Semi evaluates on ED and ES frames, while CAMUS-Full uses all available annotations. We evaluated our method using three standard medical segmentation metrics: mean Dice score (mDice), mean Hausdorff distance (mHD), and average surface distance (ASD). For the CAMUS dataset, ASD is reported in millimeters. For the EchoNet-Dynamic dataset, which lacks pixel spacing, the ASD is reported in pixels. We also employed the Wilcoxon rank-sum test to assess whether the improvement achieved by our method on the mDice metric is statistically significant. In addition, we report the statistical metrics for left ventricular ejection fraction (LVEF), including the Pearson correlation coefficient (corr) and mean bias (bias). Predicted LVEF is estimated using the Simpson's method [12].

4.2. Implementation Details

We implemented our method in PyTorch, using the pre-trained ResNet-50 backbone for robust weight initialization. Our modules were initialized following the "Kaiming" strategy [15]. All models were trained for 50 epochs using a polynomial learning-rate decay with an initial rate of 1×10^{-4} and a power of 0.9. We set the batch size to 4 and used the Adam optimizer [20] to accelerate convergence. For both CAMUS and EchoNet-Dynamic datasets, videos were uniformly sampled to 10 frames, with image resolutions standardized to 256×256 and 128×128. We cropped each sequence such that the ED frame appears first and the ES frame last. Each dataset was split into training, validation, and test subsets in a 7:1:2 ratio. During the training phase, we applied gamma correction, random scaling, random rotation, and random contrast adjustments for data augmentation, each with a probability of 0.5.

4.3. Comparison with State-of-the-art Methods

To demonstrate the effectiveness of our method, we extensively selected SOTA methods from different types, including two natural video object segmentation methods: Cutie [8], VideoMamba [25] and six echocardiography video segmentation methods: CLAS [39] and TCS [36] based on pseudo-labeling method, PKEchoNet [40], DSA [26], SAM-based improved MemSAM [10], and Mamba-based P-

Table 1. Statistical comparison with state-of-the-art methods on the CAMUS-Semi and EchoNet-Dynamic datasets. EchoForge (Full) was evaluated on the CAMUS-Full dataset.

Method	Year	CAMUS-Semi						EchoNet-Dynamic					
		mDice	P-value	mHD	ASD	corr	bias	mDice	P-value	mHD	ASD	corr	bias
Cutie	2024	88.91	0.008	5.98	2.04	0.612	8.63	87.75	0.006	4.86	1.86	0.592	9.26
VideoMamba	2024	90.13	0.015	5.43	1.80	0.653	10.24	88.78	0.018	4.57	1.52	0.624	8.68
CLAS	2020	91.46	0.023	5.05	1.52	0.795	2.03	91.05	0.036	4.10	1.36	0.788	2.85
TCS	2023	92.15	0.020	4.21	1.44	0.814	8.79	91.76	0.024	3.75	1.33	0.800	-1.22
PKEchoNet	2023	93.58	0.039	3.52	1.33	0.878	-0.31	92.64	0.031	3.32	1.21	0.843	-1.81
DSA	2024	93.65	0.041	3.45	1.25	0.891	0.52	92.75	0.028	3.22	1.15	0.871	-0.63
MemSAM	2024	93.26	0.034	4.04	1.49	0.788	4.78	92.42	0.044	3.41	1.29	0.765	4.52
P-Mamba	2024	92.93	0.031	3.84	1.38	0.853	0.64	92.14	0.026	3.58	1.30	0.823	-1.32
EchoForge	-	**94.89**	-	**3.12**	**1.18**	**0.913**	**0.23**	**93.63**	-	**3.05**	**1.02**	**0.887**	**-0.51**
EchoForge (Full)	-	94.36	-	3.26	1.22	-	-	-	-	-	-	-	-

Table 2. Efficiency comparison with the state-of-the-art methods using a single RTX 3090 GPU at 256×256 resolution.

Method	mDice	Params	FLOPs	FPS
Cutie	89.56	42M	221G	42
VideoMamba	90.13	13M	**24G**	**291**
CLAS	91.46	21M	1138G	78
TCS	92.15	156M	285G	32
PKEchoNet	93.58	41M	162G	236
DSA	93.65	**10M**	183G	81
MemSAM	93.26	257M	565G	13
P-Mamba	92.93	53M	25G	124
EchoForge	**94.89**	67M	125G	46

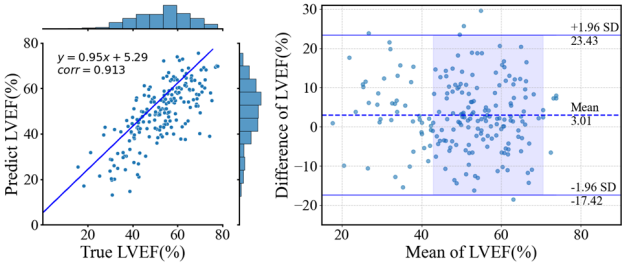


Figure 4. Linear Regression and Bland-Altman plots for the clinical metric LVEF on the CAMUS-Semi test set.

Mamba [42]. As shown in Table 1, EchoForge outperforms all SOTA methods across standard metrics. Comparing the CAMUS-Semi and CAMUS-Full datasets, the full-sequence evaluation reveals only a 0.5% reduction in mDice. This demonstrates its effectiveness in maintaining temporal consistency across the entire cardiac cycle while significantly reducing annotation costs. Additionally, the P-values of our method compared to other SOTA methods on mDice are all less than 0.05, confirming that our superior performance is statistically significant.

Table 3. Ablation study on the proposed modules on the CAMUS-Semi dataset.

Method	ASA		CPR	mDice	mHD	ASD
	TSF	ARC				
I				88.52	6.32	2.15
II	✓			92.36	4.02	1.60
III	✓	✓		93.43	3.38	1.34
IV	✓	✓	✓	**94.89**	**3.12**	**1.18**

Furthermore, Table 1 and Figure 4 show that our method achieves a stronger correlation between predicted and ground truth ejection fractions, indicating more accurate functional assessment. As shown in Figure 5, we visualized some cases at different difficulties, which further demonstrated that even in challenging cases our model produces more precise delineations, owing to the ASA module's spatiotemporal feature enhancement and and the CPR module's regeneration of high-quality pseudo-labels.

We also evaluated the computational efficiency by comparing the number of parameters (Params), floating point operations (FLOPs) and frames per second (FPS) during inference, and clearly demonstrated the trade-off between segmentation performance and efficiency of different methods. As shown in Table 2, although the VideoMamba has fewer parameters and higher FPS, it struggles in complex echocardiography video scenarios due to its limited use of

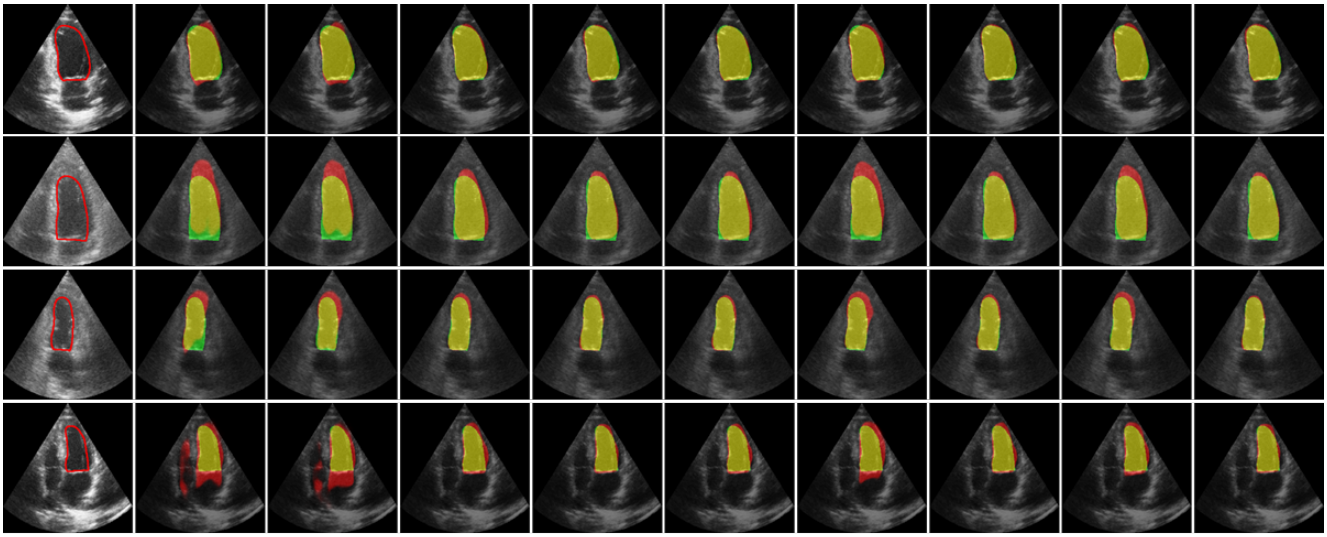


Figure 5. Visual comparison with state-of-the-art methods on the CAMUS-Semi test set. Green, red, and yellow regions represent the ground truth, prediction, and overlapping regions, respectively.

Table 4. Ablation study on varying the number of anchors on the CAMUS-Semi dataset.

Number	mDice	mHD	corr	FPS
1	94.52	3.20	90.59	92
2	94.89	3.12	91.25	46
3	94.96	3.10	91.34	35
4	94.91	3.11	91.28	23

inter-frame information. In contrast, Echoforge achieves a better balance between accuracy and efficiency. Benefiting from the ASA and CPR modules, it maintains a real-time inference speed of 46 FPS, which fully meets clinical application requirements.

4.4. Ablation Studies

We conducted ablation studies on the CAMUS-Semi dataset to assess the contribution of each component to our method's performance.

4.4.1. Effectiveness of ASA

To quantify the contributions of our two Anchor Semantic Awareness (ASA) submodules, we performed ablation studies by incrementally adding Temporal Semantic Fusion (TSF) and Anchor Recalibration (ARC) to our baseline. As shown in Table 3, incorporating TSF alone results in a notable 3.8% gain in mDice, demonstrating its effectiveness in propagating accurate anatomical cues. Adding the ARC module further improves performance, confirming its ability to suppress speckle noise. Crucially, when TSF and ARC are combined, our model achieves optimal performance across all metrics, validating the complementary synergy between semantic alignment and anchor recalibration. As shown in Figure 6, visualizations of intermediate feature maps further reveal that TSF sharpens chamber boundaries, while ARC selectively attenuates scattered artifacts, resulting in clearer and more stable representations for downstream decoding.

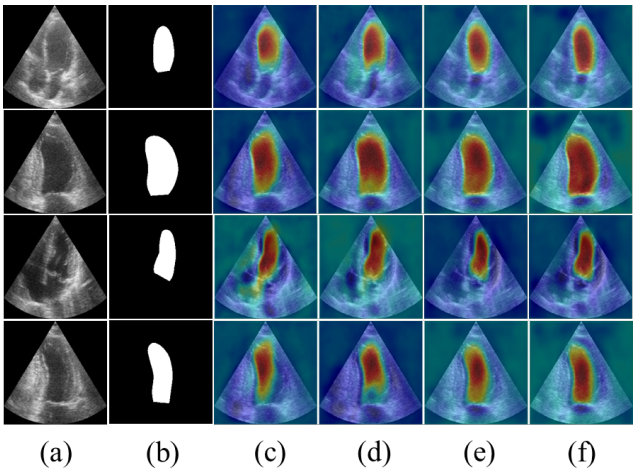


Figure 6. Qualitative ablation of different components on the CAMUS-Semi dataset. (a) Input image. (b) Ground truth. (c)-(f) Corresponding to Method (I)-(IV) in Table 3, respectively.

Furthermore, we conducted ablation studies on the internal parameters of ARC. As shown in Table 4, we experimentally determined the number of anchors, the results show

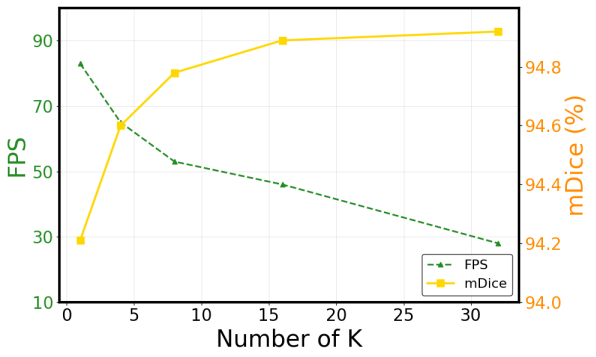


Figure 7. Performance across different values of the parameter K on the CAMUS-Semi dataset.

Table 5. Ablation study on the CPR components on the CAMUS-Semi dataset.

Method	mDice	mHD	corr	FPS
Baseline (w/o CPR)	93.43	3.38	87.15	**62**
+ channel attention	94.54	3.19	90.63	54
+ FlameRise (Ours)	**94.89**	**3.12**	**91.25**	46

that using two anchors can achieve the best performance balance. For the parameter K of KNN, Figure 7 illustrates the performance trend across varying values. Based on these results, we set $K = 16$ as the default, striking an optimal balance between performance and computational overhead.

4.4.2. Effectiveness of CPR

To demonstrate the effectiveness of Continuous Pseudo-label Reforging (CPR) in refining pseudo-label quality, we conducted further ablation experiments. As shown in Table 3, incorporating CPR yields a significant performance boost, indicating its ability to effectively suppress pseudo-label noise while enhancing the representation of critical anatomical structures. Visualizations in Figure 6 further show that masks generated with CPR exhibit improved coherence and align more closely with ground truth annotations at the left ventricular boundaries.

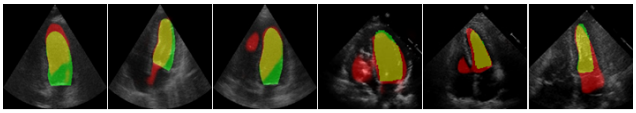
Furthermore, as shown in Table 5, we performed a stepwise ablation of the CPR submodules, demonstrating the individual contribution of each component. Additionally, to validate the stability of CPR, we compared it against alternative pseudo-labeling strategies. As demonstrated in Table 6, CPR achieves superior performance and efficiency compared to other widely used pseudo-labeling methods.

4.5. Disscussions and Limitations

Although our experiments focus primarily on sparsely annotated echocardiography video segmentation, the core mechanisms of ASA and CPR hold great potential for broader semi-supervised video segmentation tasks. Nevertheless, our method still encounters challenges in extreme scenarios. As illustrated in Figure 8, in cases with exceptionally low image contrast or severe acoustic shadowing where speckle noise completely obscures anatomical boundaries, the model struggles to recover reliable features. In such instances, even experienced cardiologists find it difficult to discern the left ventricular contours.

Table 6. Ablation study comparing various pseudo-labeling methods on the CAMUS-Semi test set.

Method		mDice	mHD	corr	FPS
ASA	-	93.43	3.38	87.15	**62**
	+CLAS	93.86	3.38	89.42	41
	+TCS	94.03	3.34	89.94	38
	+CPR	**94.89**	**3.12**	**91.25**	46



(a) (b) (c) (d) (e) (f)

Figure 8. Visualization of failure cases from the CAMUS-Semi (a-c) and EchoNet-Dynamic (d-f) test sets.

5. Conclusion

In this paper, we propose a novel semi-supervised echocardiography video segmentation model named EchoForge. Its core components include an Anchor Semantic Awareness (ASA) module and a Continuous Pseudo-label Reforging (CPR) module. ASA effectively reduces the uncertainty in noisy ultrasound segmentation by recalibrating ambiguous features through learnable anchors, and propagates structural cues across frames through semantic prototypes to improve contour accuracy. Finally, CPR continuously reforges pseudo-labels based on the channel-level reconstruction of key annotation features. Our method achieves state-of-the-art performance on the CAMUS and EchoNet-Dynamic benchmarks, showcasing strong generalization capabilities while maintaining real-time performance.

Acknowledgments

This work was supported partly by National Natural Science Foundation of China (No. 62273241), Natural Science Foundation of Guangdong Province, China (No. 2024A1515011946), the Shenzhen Research Foundation for Basic Research, China (No. JCYJ20250604181940054), and the grant under Hong Kong RGC Collaborative Research Fund (project no C5055-24G).

References

[1] Maria Antico, Fumio Sasazawa, Liao Wu, Anjali Jaiprakash, Jonathan Roberts, Ross Crawford, Ajay K Pandey, and Davide Fontanarosa. Ultrasound guidance in minimally invasive robotic procedures. *Medical image analysis*, 54:149–167, 2019. 1

[2] Navchetan Awasthi, Lars Vermeer, Louis S Fixsen, Richard GP Lopata, and Josien PW Pluim. Lvnet: Lightweight model for left ventricle segmentation for short axis views in echocardiographic imaging. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 69(6): 2115–2128, 2022. 2

[3] Maksym Bekuzarov, Ariana Bermudez, Joon-Young Lee, and Hao Li. Xmem++: Production-level video segmentation from few annotated frames. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 635–644, 2023. 2

[4] Nagashettappa Biradar, Mohan Lal Dewal, and Manoj Kumar Rohit. Speckle noise reduction in b-mode echocardiographic images: A comparison. *IETE Technical Review*, 32(6):435–453, 2015. 1

[5] Chen Chen, Chen Qin, Huaqi Qiu, Giacomo Tarroni, Jinming Duan, Wenjia Bai, and Daniel Rueckert. Deep learning for cardiac image segmentation: a review. *Frontiers in cardiovascular medicine*, 7:508599, 2020. 1

[6] Gongping Chen, Lei Li, Yu Dai, Jianxun Zhang, and Moi Hoon Yap. Aau-net: an adaptive attention u-net for breast lesions segmentation in ultrasound images. *IEEE Transactions on Medical Imaging*, 42(5):1289–1300, 2022. 2

[7] Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *European Conference on Computer Vision*, pages 640–658. Springer, 2022. 2

[8] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Joon-Young Lee, and Alexander Schwing. Putting the object back into video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3151–3161, 2024. 2, 5

[9] Junlong Cheng, Jin Ye, Zhongying Deng, Jianpin Chen, Tianbin Li, Haoyu Wang, Yanzhou Su, Ziyan Huang, Jilong Chen, Lei Jiang, et al. Sam-med2d. *arXiv preprint arXiv:2308.16184*, 2023. 2

[10] Xiaolong Deng, Huisi Wu, Runhao Zeng, and Jing Qin. Memsam: Taming segment anything model for echocardiography video segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9622–9631, 2024. 5

[11] Jiaqing Fan, Bo Liu, Kaihua Zhang, and Qingshan Liu. Semi-supervised video object segmentation via learning object-aware global-local correspondence. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12):8153–8164, 2021. 2

[12] EDWARD D Folland, ALFRED F Parisi, PAUL F Moynihan, D Ray Jones, CHARLES L Feldman, and DONALD E Tow. Assessment of left ventricular ejection fraction and volumes by real-time, two-dimensional echocardiography. a comparison of cineangiographic and radionuclide techniques. *Circulation*, 60(4):760–766, 1979. 5

[13] Raghudeep Gadde, Varun Jampani, and Peter V Gehler. Semantic video cnns through representation warping. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4453–4462, 2017. 2

[14] Libao Guo, Baiying Lei, Weiling Chen, Jie Du, Alejandro F Frangi, Jing Qin, Cheng Zhao, Pengpeng Shi, Bei Xia, and Tianfu Wang. Dual attention enhancement feature fusion network for segmentation and quantitative analysis of paediatric echocardiography. *Medical Image Analysis*, 71:102042, 2021. 2

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015. 5

[16] Ping Hu, Fabian Caba, Oliver Wang, Zhe Lin, Stan Sclaroff, and Federico Perazzi. Temporally distributed networks for fast video semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8818–8827, 2020. 2

[17] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017. 2

[18] Samvit Jain, Xin Wang, and Joseph E Gonzalez. Accel: A corrective fusion network for efficient semantic segmentation on video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8866–8875, 2019. 2

[19] Sekeun Kim, Kyungsang Kim, Jiang Hu, Cheng Chen, Zhiliang Lyu, Ren Hui, Sunghwan Kim, Zhengliang Liu, Aoxiao Zhong, Xiang Li, et al. Medivista-sam: Zero-shot medical video analysis with spatio-temporal sam adaptation. *arXiv preprint arXiv:2309.13539*, 6, 2023. 2

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[21] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 2

[22] Till Kroeger, Radu Timofte, Dengxin Dai, and Luc Van Gool. Fast optical flow using dense inverse search. In *European conference on computer vision*, pages 471–488. Springer, 2016. 2

[23] Sarah Leclerc, Erik Smistad, Joao Pedrosa, Andreas Østvik, Frederic Cervenansky, Florian Espinosa, Torvald Espeland, Erik Andreas Rye Berg, Pierre-Marc Jodoin, Thomas Grenier, et al. Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. *IEEE transactions on medical imaging*, 38(9):2198–2210, 2019. 2, 5

[24] Honghe Li, Yonghuai Wang, Mingjun Qu, Peng Cao, Chaolu Feng, and Jinzhu Yang. Echoefnet: Multi-task deep learning

network for automatic calculation of left ventricular ejection fraction in 2d echocardiography. *Computers in Biology and Medicine*, 156:106705, 2023. 1

[25] Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. Videomamba: State space model for efficient video understanding. In *European conference on computer vision*, pages 237–255. Springer, 2024. 5

[26] Jingyin Lin, Wende Xie, Li Kang, and Huisi Wu. Dynamic-guided spatiotemporal attention for echocardiography video segmentation. *IEEE Transactions on Medical Imaging*, 43 (11):3843–3855, 2024. 5

[27] Xian Lin, Yangyang Xiang, Li Zhang, Xin Yang, Zengqiang Yan, and Li Yu. Samus: Adapting segment anything model for clinically-friendly and generalizable ultrasound image segmentation. *arXiv preprint arXiv:2309.06824*, 4(11), 2023. 2

[28] Fei Liu, Kun Wang, Dan Liu, Xin Yang, and Jie Tian. Deep pyramid local attention neural network for cardiac structure segmentation in two-dimensional echocardiography. *Medical image analysis*, 67:101873, 2021. 1, 2

[29] Fadillah Maani, Asim Ukaye, Nada Saadi, Numan Saeed, and Mohammad Yaqub. Simlvseg: simplifying left ventricular segmentation in 2-d+ time echocardiograms with self-and weakly supervised learning. *Ultrasound in Medicine & Biology*, 50(12):1945–1954, 2024. 2

[30] David Nilsson and Cristian Sminchisescu. Semantic video segmentation by gated recurrent flow propagation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6819–6828, 2018. 2

[31] David Ouyang, Bryan He, Amirata Ghorbani, Neal Yuan, Joseph Ebinger, Curtis P Langlotz, Paul A Heidenreich, Robert A Harrington, David H Liang, Euan A Ashley, et al. Video-based ai for beat-to-beat assessment of cardiac function. *Nature*, 580(7802):252–256, 2020. 2, 5

[32] Jay N Paranjape, Nithin Gopalakrishnan Nair, Shameema Sikder, S Swaroop Vedula, and Vishal M Patel. Adaptivesam: Towards efficient tuning of sam for surgical scene segmentation. In *Annual Conference on Medical Image Understanding and Analysis*, pages 187–201. Springer, 2024. 2

[33] Esther Puyol-Antón, Bram Ruijsink, Baldeep S Sidhu, Justin Gould, Bradley Porter, Mark K Elliott, Vishal Mehta, Haotian Gu, Christopher A Rinaldi, Martin cowie, et al. Ai-enabled assessment of cardiac systolic and diastolic function from echocardiography. In *International Workshop on Advances in Simplifying Medical Ultrasound*, pages 75–85. Springer, 2022. 2

[34] Ayesha Saadia and Adnan Rashdi. A speckle noise removal method. *Circuits, Systems, and Signal Processing*, 37(6): 2639–2650, 2018. 1

[35] Songlin Shi, Palisha Alimu, Pazilai Mahemuti, Qingliang Chen, and Hao Wu. The study of echocardiography of left-ventricle segmentation combining transformer and cnn. *Available at SSRN 4184447*, 2022. 2

[36] Matteo Tafuro, Gino Jansen, and Ivana Išgum. Temporally consistent segmentations from sparsely labeled echocardiograms using image registration for pseudo-labels generation. In *International Workshop on Advances in Simplifying Medical Ultrasound*, pages 195–204. Springer, 2023. 2, 5

[37] Yuchao Wang, Haochen Wang, Yujun Shen, Jingjing Fei, Wei Li, Guoqiang Jin, Liwei Wu, Rui Zhao, and Xinyi Le. Semi-supervised semantic segmentation using unreliable pseudo-labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4248–4257, 2022. 2

[38] Ziqin Wang, Jun Xu, Li Liu, Fan Zhu, and Ling Shao. Ranet: Ranking attention network for fast video object segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3978–3987, 2019. 2

[39] Hongrong Wei, Heng Cao, Yiqin Cao, Yongjin Zhou, Wufeng Xue, Dong Ni, and Shuo Li. Temporal-consistent segmentation of echocardiography with co-learning from appearance and shape. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 623–632. Springer, 2020. 2, 5

[40] Huisi Wu, Jingyin Lin, Wende Xie, and Jing Qin. Super-efficient echocardiography video segmentation via proxy-and kernel-based semi-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2803–2811, 2023. 2, 5

[41] L. Yang, L. Qi, L. Feng, and *et al.* Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7236–7246, 2023. 2

[42] Zi Ye, Tianxiang Chen, Fangyijie Wang, Hanwei Zhang, and Lijun Zhang. P-mamba: Marrying perona malik diffusion with mamba for efficient pediatric echocardiographic left ventricular segmentation. *arXiv preprint arXiv:2402.08506*, 2024. 6

[43] Guang-Quan Zhou, Wen-Bo Zhang, Zhong-Qing Shi, Zhan-Ru Qi, Kai-Ni Wang, Hong Song, Jing Yao, and Yang Chen. Dsanet: Dual-branch shape-aware network for echocardiography segmentation in apical views. *IEEE Journal of Biomedical and Health Informatics*, 27(10):4804–4815, 2023. 2

[44] Hao Zhou, Lu Qi, Zhaoliang Wan, Hai Huang, and Xu Yang. Rgb-d co-attention network for semantic segmentation. In *Proceedings of the Asian conference on computer vision*, 2020. 2

[45] Yaqi Zhu, Changchun Xiong, Heng Zhao, and Yudong Yao. Sam-att: A prompt-free sam-related model with an attention module for automatic segmentation of the left ventricle in echocardiography. *IEEE Access*, 12:50335–50346, 2024. 2