Super-efficient Echocardiography Video Segmentation via Proxy- and Kernel-Based Semi-supervised Learning

Huisi Wu[1]*, **Jingyin Lin**[1], **Wende Xie**[1], **Jing Qin**[2]

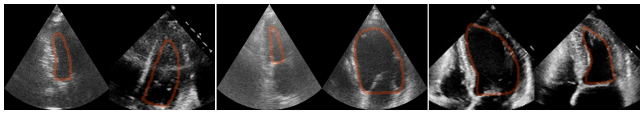
[1] College of Computer Science and Software Engineering, Shenzhen University
[2] Centre for Smart Health, The Hong Kong Polytechnic University
hswu@szu.edu.cn

Abstract

Automatic segmentation of left ventricular endocardium in echocardiography videos is critical for assessing various cardiac functions and improving the diagnosis of cardiac diseases. It is yet a challenging task due to heavy speckle noise, significant shape variability of cardiac structure, and limited labeled data. Particularly, the real-time demand in clinical practice makes this task even harder. In this paper, we propose a novel proxy- and kernel-based semi-supervised segmentation network (PKEcho-Net) to comprehensively address these challenges. We first propose a multi-scale region proxy (MRP) mechanism to model the region-wise contexts, in which a learnable region proxy with an arbitrary shape is developed in each layer of the encoder, allowing the network to identify homogeneous semantics and hence alleviate the influence of speckle noise on segmentation. To sufficiently and efficiently exploit temporal consistency, different from traditional methods which only utilize the temporal contexts of two neighboring frames via feature warping or self-attention mechanism, we formulate the semi-supervised segmentation with a group of learnable kernels, which can naturally and uniformly encode the appearances of left ventricular endocardium, as well as extracting the inter-frame contexts across the whole video to resist the fast shape variability of cardiac structures. Extensive experiments have been conducted on two famous public echocardiography video datasets, EchoNet-Dynamic and CAMUS. Our model achieves the best performance-efficiency trade-off when compared with other state-of-the-art approaches, attaining comparative accuracy with a much faster speed. The code is available at https://github.com/JingyinLin/PKEcho-Net.

Introduction

Echocardiography has been widely used for cardiac function assessment and cardiovascular disease (CVD) diagnosis (Chen et al. 2020) owing to its advantages of being real-time, economical, and noninvasive. The assessment and diagnosis are often based on the interpretation of ejection fraction (EF) and chamber volume, which depends on the accurate segmentation of key structures, such as left ventricular endocardium. In clinical practice, even for experienced clinicians, the accurate annotation of key structures in end-diastolic (ED) and end-systolic (ES) frames in echocardiography videos is time-consuming and limited by inter- and intra- observer variability (Leclerc et al. 2019). Therefore, there is a high demand for automatic segmentation approaches to accurately segment the key structures and estimate the EF from echocardiography videos.



(a) (b) (c)

Figure 1: Challenges in echocardiography video segmentation. (a) speckle noise and blurred contours. (b)-(c) the inter- and intra- sequence shape variabilities of cardiac structures.

However, it remains a challenging task due to the following reasons. First, compared to other commonly used medical imaging modalities, such as CT scans and magnetic resonance imaging (MRI), ultrasound (US) images are full of speckle noise, which makes the boundaries of the targeting objects difficult to be determined (Figure 1 (a)). Second, in inter- and intra-frames of echocardiography videos, cardiac structures often have significant shape variability (Figure 1 (b, c)). Third, due to the busy schedule of physicians, they have no time to provide pixel-level annotations across the whole video, and only ED and ES frames are manually annotated. Fourth and more importantly, to fulfill the routine clinical requirements, we have to tackle these challenges in a real-time manner; in such a case, efficiency is of great significance as accuracy in this task.

In recent years, several deep learning models have been proposed to address these challenges. The success of optical flow based methods in natural video segmentation motivated their application to echocardiographic videos (Ta et al. 2020; Wei et al. 2020). These methods employ motion field estimation to generate smooth pseudo-labels for unlabeled frames and rely on various constraints to enhance the temporal consistency of segmentation results. However, speckle noise significantly affects the quality of the estimated motion field and results in unsatisfactory performance. To deal with the influence of speckle noise, several methods (Chen et al. 2021; Wu et al. 2022a; Liu et al. 2021) exploit the inherent characteristics of 2D or 3D convolutions or local

*Corresponding author. Email: hswu@szu.edu.cn.

attention mechanisms to learn more effective features in restricted contexts for more robust temporal consistency. However, they only take the contexts of neighboring frames into account, probably due to the real-time requirements. Different from the above methods, Painchaud et al. (Painchaud et al. 2022) proposes a post-processing method to enhance temporal consistency, which corrects temporal inconsistencies via regularized shape encoding based on the segmentation results of an existing model. However, its accuracy largely depends on the segmentation model and its efficiency is also not satisfactory. On the other hand, several methods also have been proposed to develop lightweight networks to fulfill the real-time requirement. A recently proposed method (Zamzmi et al. 2022) transfers relevant knowledge for fine-tuning on top of a lightweight framework to enhance generalization and speed up convergence. But it failed to exploit temporal relationships between video frames and hence still cannot achieve satisfactory accuracy.

In this paper, we propose a novel proxy- and kernel-based semi-supervised echocardiography video segmentation to comprehensively address the above-mentioned challenges; we focus on the segmentation of left ventricular endocardium, which is essential for the assessment and diagnosis of many CVDs. To mitigate the effect of speckle noise, we propose a multi-scale region proxy (MRP) mechanism to model the region-wise contexts, in which a learnable region proxy with an arbitrary shape is developed in each layer of the encoder, helping identify homogeneous semantics and hence alleviate the influence of speckle noise on segmentation. More importantly, according to the reconstructed proxy relationships, our network is able to generate fine-grained predictions without any decoder, thereby greatly improving the inference efficiency. To fully exploit and transmit temporal contexts among labeled and unlabeled frames, we further formulate the semi-supervised segmentation using a group of learnable kernels, which is capable of naturally and uniformly encoding the appearances of left ventricular endocardium, as well as extracting the inter-frame contexts across the whole video to resist the fast shape variability of cardiac structures. Extensive experiments on two famous benchmarks, EchoNet-Dynamic (Ouyang et al. 2020) and CAMUS (Leclerc et al. 2019), demonstrate that the proposed model achieves the best performance-efficiency trade-off when compared with other state-of-the-art approaches. Our main contributions can be summarized as follows:

- We propose a novel network for semi-supervised echocardiography video segmentation which replaces the decoder by building region proxies with arbitrary shapes in each layer of the encoder, and formulates the semi-supervised segmentation with a set of learnable kernels.
- While the MRP mechanism can resist the influence of speckle noise, the learnable kernels are able to sufficiently exploit the temporal consistency for accurate segmentation; working together, they comprehensively address the challenges of this task with high efficiency.
- Our model achieves state-of-the-art results on both two famous benchmarks, attaining comparative accuracy as state-of-the-art methods with a much faster speed.

Related Works

Echocardiography Video Segmentation

Echocardiography video segmentation requires exploring the temporal consistency between video frames, and improving the segmentation accuracy of labeled frames by exploiting the information of unlabeled frames. Optical flow is widely used in video semantic segmentation tasks, several methods (Li et al. 2019; Ta et al. 2020; Wei et al. 2020; Wu et al. 2020) have attempted to use motion estimation for echocardiographic video tasks, but motion estimation is noise sensitive. To deal with the influence of speckle noise, several methods (Chen et al. 2021; Wu et al. 2022a; Liu et al. 2021) exploit the inherent characteristics of 2D or 3D convolutions or local attention mechanisms to learn more effective features in restricted contexts for more robust temporal consistency, but they only take the contexts of neighboring frames into account. Different from the above method, Painchaud et al. (Painchaud et al. 2022) proposes a post-processing method to enhance temporal consistency, but its accuracy largely depends on the segmentation model. On the other hand, several methods have been proposed to develop lightweight networks to fulfill the real-time requirements. Inspired by BiSeNet (Yu et al. 2018), Zamzmi et al. (Zamzmi et al. 2021) proposed a region-of-interest based trilateral attention network (TaNet) for real-time cardiac region segmentation. In another work (Zamzmi et al. 2022), they fine-tune TaNet by transferring relevant knowledge to enhance model generalization and speed up convergence. TransBridge (Deng et al. 2021) and LVNet (Awasthi et al. 2022) develop a lightweight echocardiography segmentation model through channel splitting. However, all four methods above fail to exploit the temporal relationships between video frames.

Dynamic Kernel

Shape-fixed receptive fields lead to static convolution kernels to learn coarse representations that are insufficient for dense labeling. To solve this problem, Dai et al. (Dai et al. 2017) proposed a deformable convolutions module to improve the modeling ability of geometric changes. Specifically, a parallel network is first used to learn the offsets, and then these offsets are added to the position of each sampling point in the convolution kernel, thereby deforming the receptive field of the model without being restricted by the regular grid. On this basis, a variety of dynamic kernels (Wu et al. 2018; He, Deng, and Qiao 2019; Zhu et al. 2019; Gao et al. 2020; Li and Chen 2021; Tian, Gao, and Peng 2022) were proposed. Different from the above methods that uses dynamic kernels to extract fine-grained features, recent studies apply dynamic kernels to directly generate segmentation predictions (Fang et al. 2021; Wang et al. 2021; Zhang et al. 2021; Li et al. 2022a; Wu et al. 2022b; Yu et al. 2022), where a set of learnable parameters are used to dynamically generate segmentation kernels for more accurate segmentation predictions. We propose a MRP module and kernel-based semi-supervised segmentation (KSS) mechanism to effectively resist the speckle noise and improve the temporal consistency respectively.

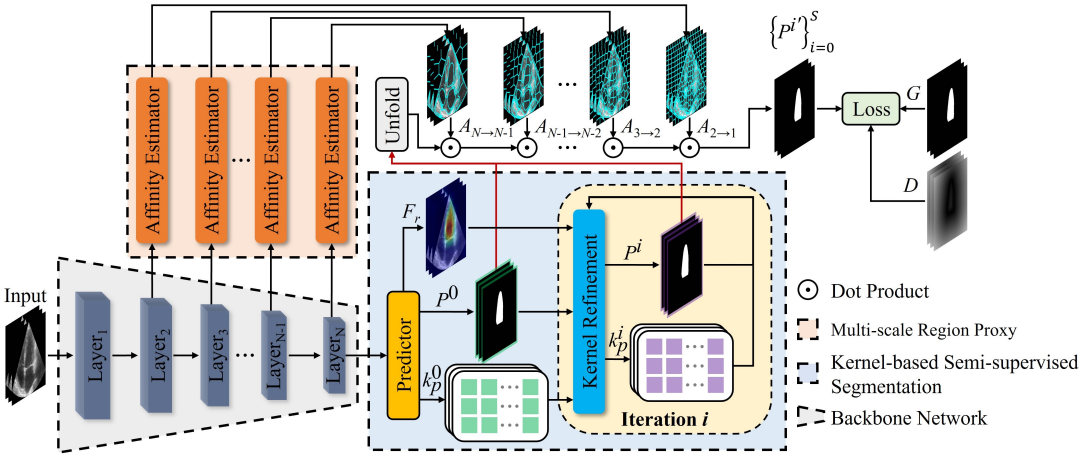


Figure 2: Overview of our PKEcho-Net, which mainly consists of an MRP module and a KSS mechanism for semi-supervised segmentation of echocardiography videos (only the first frame in ED and the last frame in ES are annotated).

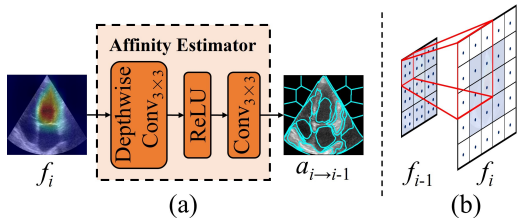


Figure 3: More details of MRP. (a) Our MRP mechanism in each layer of the encoder. (b) Region proxies are established by learning the relationships between the 3×3 neighborhood of each pixel in f_i and the corresponding regions in f_{i-1}.

Methods

Overview

The architecture of our PKEcho-Net is illustrated in Figure 2. For each frame of the input echocardiography video, we first propose an MRP module to alleviate the interference of speckle noise in segmentation. On the other hand, considering that only the first frame in ED and the last frame in ES are manually annotated in each echocardiography video, we formulate the semi-supervised echocardiography video segmentation with a group of learnable kernels, which can naturally and uniformly encode the identical left ventricular appearances and the inter-frame contexts across the whole video. Hence, our PKEcho-Net can better transmit the global temporal appearance context to resist the fast shape variability of cardiac structures during the whole video.

Multi-scale Region Proxy

The quality of features in our task may greatly degraded by massive speckle noise existing in echocardiographical images. To alleviate the damage of speckle noise, we propose an MRP module to capture the region-wise context in a more flexible manner. Different from traditional methods, which usually relied on a rigid tessellation on the input image to model region-wise context with constrained receptive fields (Liu et al. 2021; Wu et al. 2022a), we employ multi-scale learnable region proxies in each layer of the encoder, which can have arbitrary shapes and denote the homogeneous semantics to resist the speckle noise, as shown in Figure 3. Unlike the single region proxy proposed by (Zhang, Pang, and Lu 2022), our MRP module consists of $N-1$ affinity estimators, where N is the layer number of the backbone network. In this regard, we can establish the association between the pixels in the current layer and the pixels in the corresponding homogeneous region of the previous layer, thereby further improving the anti-noise ability.

Specifically, we employ two 3×3 convolutional layers with an inner ReLU activation to construct the affinity estimator between two neighboring feature maps, as shown in Figure 3 (a). To further reduce computation cost, we adopt depth-wise convolution (Howard et al. 2017) in the first layer. Suppose that an output feature map of the i-th layer in the backbone network is denoted as $f_i \in \mathbb{R}^{C\times\frac{H}{h}\times\frac{W}{w}}$, where $i=2,\ldots,N$, C is the number of channels, $H\times W$ is the resolution of f_{i-1}, and h or w is the downsampling rate. To alleviate the damage of heavy speckle noise, we can model region-wise context based on the appearance-level relationships between the neighborhood of each pixel in f_i and $h\times w$ non-overlapping regions in f_{i-1}, as shown in Figure 3 (b). For each pixel in the corresponding $h\times w$ region of f_{i-1}, we need define a $M\times M$ neighbor pixels to learn a proxy pixel with homogeneous semantics in f_i. In our experiments, we can set $M=3$ to obtain a good balance between accuracy and efficiency. Thus, we adopt 3×3 instead of 1×1 convolutional layers to build the affinity estimator, which can better establish appropriate proxy relationships to resist heavy speckle noise. Finally, we can obtain an affinity map $a_{i\to i-1}\in\mathbb{R}^{M\times M\times(H\,h)\times(W\,w)}$ to describe the association of pixels in f_i with the corresponding proxy region in f_{i-1}.

Kernel-based Semi-supervised Segmentation

Our task is a typical, yet special, semi-supervised segmentation task, where only the first frame in ED and the last frame

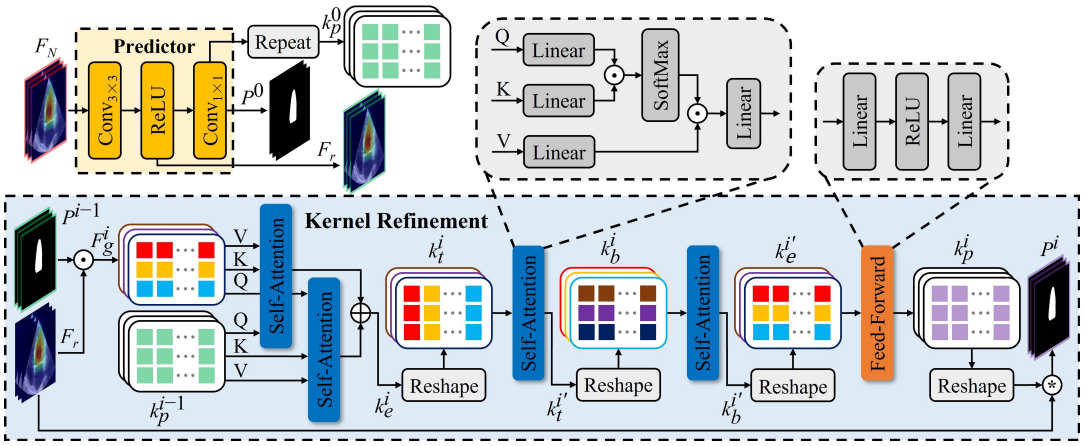


Figure 4: Illustration of KSS. We iteratively train a group of learnable kernels to uniformly encode the identical left ventricular appearances and the inter-frame contexts across the whole video, which is also much more compatible and memory efficient.

in ES are annotated in each echocardiography video. To fully exploit and transmit the temporal appearance contexts among labeled frames and unlabelled frames, we propose a KSS mechanism to uniformly encode the identical left ventricular appearances and the inter-frame contexts across the whole video by a group of learnable kernels. Specifically, we initialize the group of kernel from the features extracted by the backbone network through a predictor, then introduce hybrid attention and self-attention over temporal and different videos to further refine the kernels for better mask predictions. Compared with traditional methods which only exploit temporal context of two neighboring frames through feature warping (Ta et al. 2020; Wei et al. 2020) or self-attention-based mean teacher semi-supervised architecture (Liu et al. 2021; Wu et al. 2022a), our PKEcho-Net is able to better transmit the global temporal appearance context to resist the fast shape variability of cardiac structures during the whole video. More importantly, the KSS mechanism is simply implemented with a group of kernel, which is also much more compatible and memory efficient and can be easily equipped in most deep learning architectures.

As shown in Figure 4, given the feature maps sequence $F_N \in \mathbb{R}^{(BT)\times C\times H\times W}$, $F_N = \{f_N^l\}_{l=1}^T$, extracted from the last layer in the backbone network, we first apply a predictor to reduce the channel and output a coarse mask prediction sequence $P^0 \in \mathbb{R}^{(BT)\times c\times H\times W}$, where B is the batch size and c is the number of classes. The predictor consists of two convolutional layers with an inner ReLU activation, including a 3×3 convolutional layer for the channel reduction and a 1×1 convolutional layer for the mask prediction.

During the iterations for the kernel refinement, we first initialize the group of kernel with 1×1 convolutional kernel $k_p^0 \in \mathbb{R}^{(BT)\times c\times 1\times d}$ with expanded dimensions, where d is the number of channels after reduction. Suppose that the feature map after channel reduction is $F_r \in \mathbb{R}^{(BT)\times d\times H\times W}$. To refine the kernel based on the intra-frame appearance context, we apply a dot product between the feature maps sequence F_r and the P^0 to group the pixels with similar semantics. Thus, we can formulate the grouped feature maps sequence in the i^{th} iteration as

$$F_g^i = \sum_x^H \sum_y^W P^{i-1}(x,y)\cdot F_r(x,y) \tag{1}$$

where $F_g^i \in \mathbb{R}^{(BT)\times c\times 1\times d}$ is the grouped feature maps.

Given the grouped feature maps sequence, we further adopt a hybrid attention mechanism (Li et al. 2022a) to learn more effective distinct cues and appearance semantics in each frame. Specifically, the implementation of hybrid attention is similar to vanilla self-attention ($\mathcal{F}_{vsa}$) (Vaswani et al. 2017), where only the query (Q), key (K) and value (V) are denoted with different inputs. Based on the bidirectional attention between feature map and learned kernel, we can obtain a kernel embedding $k_e^i \in \mathbb{R}^{(BT)\times c\times 1\times d}$ written as

$$\begin{aligned} k_e^i =& (\mathcal{F}_{vsa}(F_g^i, k_p^{i-1}, k_p^{i-1}) + F_g^i) + \\ & (\mathcal{F}_{vsa}(k_p^{i-1}, F_g^i, F_g^i) + k_p^{i-1}) \end{aligned} \tag{2}$$

$$\mathcal{F}_{vsa}(Q,K,V) = \psi_4\left(\sigma\left(\frac{\psi_1(Q)\psi_2(K)^T}{\sqrt{d}}\right)\psi_3(V)\right) \tag{3}$$

where $\psi_j, j = 1,\ldots,4$ are different linear layers and σ is a Softmax function. We employ LayerNorm (LN) after the residual connection between the output of $\mathcal{F}_{vsa}$ and its input query, and the same is true for the following.

In addition, we also use vanilla self-attention to model pair-wise relationships in temporal dimension to enhance the temporal consistency among different kernel embeddings during the whole video. By restoring the dimension order of the kernel embeddings with a reshape operator, we can adopt a feed-forward network ($\mathcal{F}_{ffn}$) (Vaswani et al. 2017) to predict a new kernel k_p^i for segmentation, where the dimension of the inner-layer should be the same as the input. The above process can be written as

$$\begin{aligned} k_t^{i'} &= \mathcal{F}_{vsa}(k_t^i, k_t^i, k_t^i) + k_t^i, k_t^i = \phi_1(k_e^i) \\ k_b^{i'} &= \mathcal{F}_{vsa}(k_b^i, k_b^i, k_b^i) + k_b^i, k_b^i = \phi_2(k_t^{i'}) \\ k_p^i &= \mathcal{F}_{ffn}(k_e^{i'}) + k_e^{i'}, k_e^{i'} = \phi_3(k_b^{i'}) \end{aligned} \tag{4}$$

Method	Backbone	CAMUS						EchoNet-Dynamic					
		DSC	P-value	HD	ASD	corr	bias±std	DSC	P-value	HD	ASD	corr	bias±std
VKNet	ResNet-50	92.8	0.036	4.2	1.4	83.3	-2.6±8.7	92.2	0.031	3.3	1.2	79.1	1.8±9.5
HITF	ResNet-50	92.7	0.033	4.3	1.4	82.7	-1.8±9.1	92.2	0.028	3.3	1.2	78.3	2.2±10.1
CLAS	UNet-3D	92.1	0.042	4.8	1.6	81.0	-3.8±8.8	92.1	0.024	3.3	1.2	80.8	2.9±9.2
PLANet	ResNet-34	92.9	0.026	4.1	1.4	84.7	0.5±8.3	91.7	0.020	3.4	1.3	76.5	-1.2±10.7
TaNet	ResNet-18	92.5	0.020	4.6	1.5	83.4	-2.4±8.5	92.0	0.045	3.3	1.2	78.8	1.2±9.7
LVNet	-	91.7	0.047	5.0	1.6	75.7	-1.1±10.7	92.2	0.022	3.3	1.2	81.2	-1.2±9.2
SSCF	ResNet-18	93.3	0.031	4.0	1.3	85.8	0.5±8.2	92.3	0.023	3.3	1.2	83.3	-0.9 ± 8.6
Ours	ResNet-18	93.7	0.029	3.6	1.2	87.5	-0.7±7.4	92.7	0.034	3.1	1.1	84.2	-1.8 ± 8.3
	ResNet-34	93.9	0.038	3.5	1.2	86.0	-0.6±8.0	92.8	0.037	3.1	1.1	85.3	**-0.8±8.0**
	ResNet-50	**94.2**	-	**3.4**	**1.1**	**89.9**	**-0.3 ± 6.7**	**93.1**	-	**3.1**	**1.1**	**87.4**	-1.6±7.4

Table 1: Statistical comparison with state-of-the-art methods on the CAMUS and EchoNet-Dynamic test sets.

Method	Flops	Params	FPS
VKNet (Li et al. 2022b)	124G	29.0M	216
HITF (Li et al. 2022a)	718G	41.0M	91
CLAS (Wei et al. 2020)	1125G	19.1M	82
PLANet (Liu et al. 2021)	721G	37.8M	142
TaNet (Zamzmi et al. 2021)	278G	26.3M	155
LVNet (Awasthi et al. 2022)	**32G**	**0.4M**	50
SSCF (Wu et al. 2022a)	625G	57.7M	62
Ours (ResNet-18)	42G	11.7M	**428**
Ours (ResNet-34)	80G	21.8M	312
Ours (ResNet-50)	109G	25.7M	258

Table 2: Efficiency comparison with the state-of-the-art methods on one RTX 3090 GPU at 320×320 resolution.

where $\phi_j, j = 1, \ldots, 3$ are different dimension transformations, $k_t^i \in \mathbb{R}^{(B\,c)\times T\times d}$, $k_b^i \in \mathbb{R}^{(T\,c)\times B\times d}$, and $k_e^{i\prime} \in \mathbb{R}^{(B\,T)\times c\times 1\times d}$. We also employ LN after the residual connection between the output of $\mathcal{F}_{ffn}$ and its input query.

Finally, by applying a convolution between the feature maps F_r and the kernel prediction k_p^i, we can generate a finer mask prediction $P^i = k_p^i * F_r$ for the next iteration of kernel refinement and loss function computation. Before calculating the loss, we can gradually recover the size of P^i through the affinity maps sequence $\{A_{n\to n-1}\}_{n=2}^{N}$, $A_{n\to n-1} = \{a_{n\to n-1}^l\}_{l=1}^{T}$, estimated with our MRP mechanism, which can be written as

$$P^{i\prime} = \sum_{j=1}^{M\times M} \left(P^i(j) \cdot \{A_{n\to n-1}(j)\}_{n=2}^{N}\right) \tag{5}$$

where $M \times M$ is the number of neighbor pixels to learn a proxy pixel with homogeneous semantics.

Loss Functions

We adopt a Dice loss $\mathcal{L}_{dice}$ (Milletari, Navab, and Ahmadi 2016) to supervise our PKEcho-Net. To further improve the anti-noise ability in handling the finely detailed boundaries of LV, we also introduce a boundary loss $\mathcal{L}_b$ (Kervadec et al. 2021) in training our PKEcho-Net, which can adaptively guide the inner segmented region in the echocardiography by measuring the distance-to-boundary information on the basis of $\mathcal{L}_{dice}$. Moreover, we supervise each mask prediction $P^{i\prime}, i = 0, \ldots, S$, S is the number of iterations, to ensure that kernel refinement iterations do not deviate right direction during the kernel-based semi-supervised segmentation. Finally, the total loss $\mathcal{L}_{total}$ of our method can be written as

$$\mathcal{L}_{total} = \sum_{i=0}^{S} \Big(\mathcal{L}_{dice}(P^{i\prime}, G) + \mathcal{L}_b(P^{i\prime}, D)\Big) \tag{6}$$

where G is the ground truth, and D is the distance map precomputed from G.

Method	Interval	DSC	HD	ASD
ResNet-50	-	92.3	4.5	1.6
ResNet-50 + VRP	4	92.7	4.3	1.4
ResNet-50 + MRP	4	93.1	4.1	**1.2**
	2	93.3	3.9	1.3
	1	**93.5**	**3.7**	1.3

Table 3: Ablation study of MRP on the CAMUS test set.

Experiments

Datasets and Evaluation Metrics

We evaluated our method on two public echocardiography video datasets: EchoNet-Dynamic (Ouyang et al. 2020) and CAMUS (Leclerc et al. 2019) datasets.

- **EchoNet-Dynamic** contains 10,030 apical-4-chamber echocardiography videos, each of which was cropped and masked to remove text and information outside the scanning sector. All videos are downsampled to a normalized resolution of 112×112.
- **CAMUS** contains 450 cases from both healthy volunteers and patients, each of which further includes an apical-2-chamber and an apical-4-chamber echocardiography video. Thus, CAMUS has 900 videos with an av-

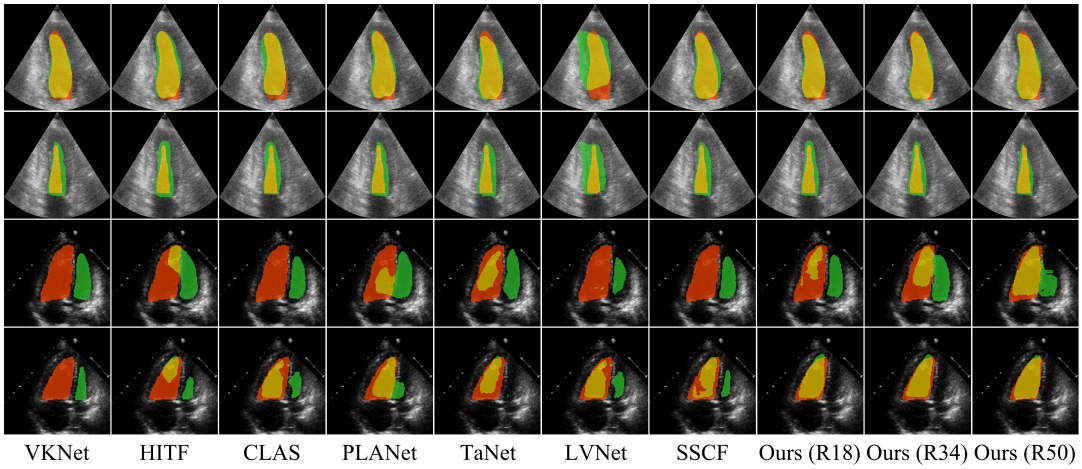


Figure 5: Visual comparison with different state-of-the-art methods on the CAMUS and EchoNet-Dynamic test sets. Red, green, and yellow regions represent the ground truth, prediction, and their overlapping regions, respectively.

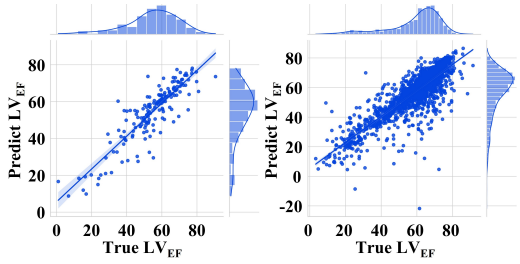


Figure 6: Correlation graphs for the clinical metrics on the CAMUS (left) and EchoNet-Dynamic (right) test sets.

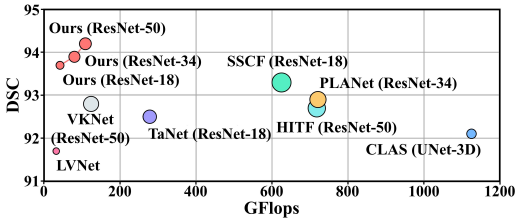


Figure 7: Performance vs. efficiency on the CAMUS test set.

erage resolution of 984×600, where 19% of the cases have poor quality with lower contrast and heavier noise.

In our experiments, we employed four widely-used metrics to evaluate the echocardiography video segmentation: Dice Similarity Coefficient (DSC), Hausdorff Distance-95% (HD), and Average Surface Distance (ASD), and a clinical metric, Left Ventricular Ejection Fraction (LV_{EF}).

Implementation Details

We implemented our method with the PyTorch framework and adopted the pre-trained ResNet-50 (He et al. 2016) as the backbone network to obtain a relatively better initialization. For the proposed MRP and KSS modules, we adopted the "Kaiming" strategy (He et al. 2015) to initialize the parameters. We trained our model for 50 epochs with a poly strategy, where the learning rate is multiplied by $(1 - \frac{iter}{iter_{max}})^{0.9}$ for each iteration with an initial learning rate of 1e-3 for all experiments. We set batchsize $= 8$ and an Adam optimizer (Kingma and Ba 2014) is also used to accelerate the convergence. By unifying the image resolution of the EchoNet-Dynamic and CAMUS to 128×128 and 320×320, each video in both datasets is also equidistantly reshaped into 10 frames. We split the training set, validation set, and test set with a ratio of 7:1:2, where four kinds of data augmentations are used to enrich the video data diversity for training, including horizontal flipping, randomly turning the brightness or contrast in a range of [-15%, +15%] and random rotations in [-20, 20] degrees.

Comparison with State-of-the-art Methods

We compared our method with seven state-of-the-art methods on both the EchoNet-Dynamic and CAMUS datasets, including two most relevant natural video segmentation methods (VKNet (Li et al. 2022b) and HITF (Li et al. 2022a)) and five latest echocardiography video segmentation methods (CLAS (Wei et al. 2020), PLANet (Liu et al. 2021), TaNet (Zamzmi et al. 2021), LVNet (Awasthi et al. 2022) and SSCF (Wu et al. 2022a)). As shown in Table 1, our method generally outperforms other competitors with the highest performances in all metrics on both datasets. We further performed the Wilcoxon rank-sum test of DSC, and our method (ResNet-50 based) has a statistical improvement of DSC at the 5% level (all P-values are less than 0.05). From the visual comparisons of typical challenging test cases from both datasets, we clearly observe that even for the complex regions near the LV endocardium, our results are still the closest to the ground truth, as shown in Figure 5.

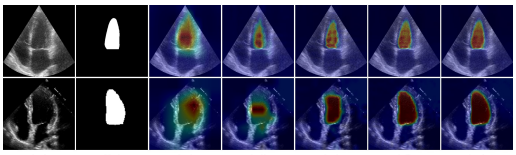
Moreover, we conducted a statistical comparison using the clinical metrics on both datasets, including the Pearson correlation coefficient (corr), mean bias (bias), and standard deviations (std) of LV_{EF} between ground truth and the prediction, where LV_{EF} is calculated with Simpson's rule (Fol-

Method	H.A.	T.A.	DSC	HD	ASD
ResNet-50	-	-	92.3	4.5	1.6
ResNet-50 + KSS			92.8	4.2	1.4
	✓		93.2	3.9	1.3
	✓	✓	**93.6**	**3.6**	**1.2**

Table 4: Ablation study of KSS on the CAMUS test set.

Iteration #	DSC	HD	ASD
1	93.7	3.6	1.2
2	94.0	3.5	1.2
3	**94.2**	**3.4**	**1.1**
4	93.9	3.5	1.2

Table 5: Ablation study of the iteration # on CAMUS.



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

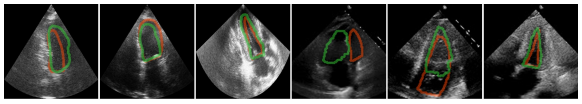
Figure 8: Visual comparison of feature maps restored from different affinity maps. (a) Input image. (b) Ground truth. (c) feature map extracted in the backbone. (d-g) feature maps gradually restored to input image size using affinity maps.

land et al. 1979). As shown in Table 1, our method achieve a higher correlation and lower deviation on LV_{EF} than all other methods, which further improves the confidence to analyze and evaluate cardiac function through echocardiographic segmentation. In Figure 6, we present the correlation graphs of the proposed method (ResNet-50 based) on both two datasets, where we can also observe a good consistency between the ground truth and predicted clinical indices.

On the other hand, we also demonstrated the advantages of our proposed method in efficiency by calculating the number of parameters (Params) and floating point operations (Flops), and frames per second (FPS) during inference of different networks. As shown in Table 2, we achieve the fastest inference speed, which benefits from the MRP module in saving a heavy decoder and the KSS mechanism in kernel-level operation. Although LVNet exhibits the least amount of computation and parameters, it suffers from the worst inference speed due to excessive memory accesses. As shown in Figure 7, our method obtains a better performance-efficiency trade-off than all other competitors.

Ablation Studies

Multi-scale region proxy. To demonstrate the importance of region proxy between adjacent layers in the encoder, we composed various MRP modules by introducing affinity estimators with different interval layers. As shown in Table 3, our MRP module can improve the performance of the backbone network in all indicators, where the performance is better for the fewer number of interval layers. Moreover, by establishing semantic correlations between different scales to resist speckle noise, our MRP module still outperforms the vanilla region proxy (VRP) of (Zhang, Pang, and Lu 2022). We also visualized the feature maps with gradually recovered resolution using multi-scale affinity maps. As shown in Figure 8, the feature maps become more and more clear, indicating that our MRP module is sufficient to save the heavy decoder in echocardiography video segmentation.



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

Figure 9: Failure cases on the CAMUS (a-c) and EchoNet-Dynamic (d-f) test sets. Red and green contours denote the ground truth and our prediction, respectively.

Hybrid attention and temporal self-attention in KSS. To demonstrate the effectiveness of hybrid attention (H.A.) for kernel-feature interactions and temporal self-attention (T.A.) for modeling global temporal appearance context, we conducted ablation studies by skipping different attention mechanisms. Note that when the hybrid attention is disabled, we adopt the gating mechanism from (Zhang et al. 2021) instead. As shown in Table 4, hybrid attention outperforms the gating mechanism, because the long-term dependency learning enables the kernel to better encode distinct cues in each frame and leverage them to enhance the appearance context. Similarly, temporal self-attention also improves performance by modeling the global temporal context.

Iteration # in KSS. We also evaluated the impact of the iteration # for kernel refinements in KSS. As shown in Table 5, we can obtain the best performance after three iterations. However, if we further train the kernels, the performance may drops slightly due to the unavoidable small amount of noise carried in the over-iterations.

Discussions and Limitations

Although the above experiments only focus on the echocardiography video segmentation task, we believe our method has potential for other semi-supervised video segmentation tasks with sparse annotations. Moreover, our method still has some limitations. As shown in Figure 9, our PKEchoNet still cannot handle the videos with extremely low contrast or too heavy noise, where the LV boundaries cannot be distinguished even by an experienced cardiologist.

Conclusion

In this paper, we propose a novel and efficient semi-supervised segmentation method for echocardiography video that replaces the decoder by building region proxies with arbitrary shapes in each layer of the encoder and formulates the semi-supervised echocardiography video segmentation with a group of learnable kernels. Our method achieves state-of-the-art results on both the CAMUS and EchoNet-Dynamic test sets with the fastest inference speed.

Acknowledgments

This work was supported partly by National Natural Science Foundation of China (Nos. 61973221 and 62273241), Natural Science Foundation of Guangdong Province, China (No. 2019A1515011165), the COVID-19 Prevention Project of Guangdong Province, China (No. 2020KZDZX1174), the Major Project of the New Generation of Artificial Intelligence (No. 2018AAA0102900), and the Hong Kong Research Grant Council under General Research Fund Scheme (Project no. 15205919).

References

Awasthi, N.; Vermeer, L.; Fixsen, L. S.; Lopata, R. G.; and Pluim, J. P. 2022. 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.

Chen, C.; Qin, C.; Qiu, H.; Tarroni, G.; Duan, J.; Bai, W.; and Rueckert, D. 2020. Deep learning for cardiac image segmentation: a review. *Frontiers in Cardiovascular Medicine*, 7: 25.

Chen, Y.; Zhang, X.; Haggerty, C. M.; and Stough, J. V. 2021. Assessing the generalizability of temporally coherent echocardiography video segmentation. In *Medical Imaging 2021: Image Processing*, volume 11596, 463–469. SPIE.

Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; and Wei, Y. 2017. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 764–773.

Deng, K.; Meng, Y.; Gao, D.; Bridge, J.; Shen, Y.; Lip, G.; Zhao, Y.; and Zheng, Y. 2021. Transbridge: A lightweight transformer for left ventricle segmentation in echocardiography. In *International Workshop on Advances in Simplifying Medical Ultrasound*, 63–72. Springer.

Fang, Y.; Yang, S.; Wang, X.; Li, Y.; Fang, C.; Shan, Y.; Feng, B.; and Liu, W. 2021. Instances as queries. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6910–6919.

Folland, E.; Parisi, A.; Moynihan, P.; Jones, D. R.; Feldman, C. L.; and Tow, D. 1979. 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.

Gao, H.; Zhu, X.; Lin, S.; and Dai, J. 2020. Deformable Kernels: Adapting Effective Receptive Fields for Object Deformation. In *International Conference on Learning Representations*.

He, J.; Deng, Z.; and Qiao, Y. 2019. Dynamic multi-scale filters for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3562–3572.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, 1026–1034.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Howard, A. G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; and Adam, H. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Kervadec, H.; Bouchtiba, J.; Desrosiers, C.; Granger, E.; Dolz, J.; and Ayed, I. B. 2021. Boundary loss for highly unbalanced segmentation. *Medical image analysis*, 67: 101851.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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

Li, M.; Zhang, W.; Yang, G.; Wang, C.; Zhang, H.; Liu, H.; Zheng, W.; and Li, S. 2019. Recurrent aggregation learning for multi-view echocardiographic sequences segmentation. In *International conference on medical image computing and computer-assisted intervention*, 678–686. Springer.

Li, X.; Wang, J.; Li, X.; and Lu, Y. 2022a. Hybrid Instance-aware Temporal Fusion for Online Video Instance Segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36(2), 1429–1437.

Li, X.; Zhang, W.; Pang, J.; Chen, K.; Cheng, G.; Tong, Y.; and Loy, C. C. 2022b. Video k-net: A simple, strong, and unified baseline for video segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18847–18857.

Li, Y.; and Chen, Y. 2021. Revisiting Dynamic Convolution via Matrix Decomposition. In *International Conference on Learning Representations*.

Liu, F.; Wang, K.; Liu, D.; Yang, X.; and Tian, J. 2021. Deep pyramid local attention neural network for cardiac structure segmentation in two-dimensional echocardiography. *Medical Image Analysis*, 67: 101873.

Milletari, F.; Navab, N.; and Ahmadi, S.-A. 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, 565–571. IEEE.

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

Painchaud, N.; Duchateau, N.; Bernard, O.; and Jodoin, P.-M. 2022. Echocardiography segmentation with enforced temporal consistency. *IEEE Transactions on Medical Imaging*.

Ta, K.; Ahn, S. S.; Stendahl, J. C.; Sinusas, A. J.; and Duncan, J. S. 2020. A semi-supervised joint network for simultaneous left ventricular motion tracking and segmentation in

4D echocardiography. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 468–477. Springer.

Tian, H.; Gao, P.; and Peng, X. 2022. Video Frame Interpolation Based on Deformable Kernel Region. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, H.; Zhu, Y.; Adam, H.; Yuille, A.; and Chen, L.-C. 2021. Max-deeplab: End-to-end panoptic segmentation with mask transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5463–5474.

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

Wu, H.; Liu, J.; Xiao, F.; Wen, Z.; Cheng, L.; and Qin, J. 2022a. Semi-supervised segmentation of echocardiography videos via noise-resilient spatiotemporal semantic calibration and fusion. *Medical Image Analysis*, 78: 102397.

Wu, J.; Li, D.; Yang, Y.; Bajaj, C.; and Ji, X. 2018. Dynamic filtering with large sampling field for convnets. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 185–200.

Wu, J.; Yarram, S.; Liang, H.; Lan, T.; Yuan, J.; Eledath, J.; and Medioni, G. 2022b. Efficient video instance segmentation via tracklet query and proposal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 959–968.

Wu, P.; Huang, Q.; Yi, J.; Qu, H.; Ye, M.; Axel, L.; and Metaxas, D. 2020. Cardiac MR image sequence segmentation with temporal motion encoding. In *European Conference on Computer Vision*, 298–309. Springer.

Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; and Sang, N. 2018. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, 325–341.

Yu, X.; Shi, D.; Wei, X.; Ren, Y.; Ye, T.; and Tan, W. 2022. SOIT: Segmenting Objects with Instance-Aware Transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36(3), 3188–3196.

Zamzmi, G.; Rajaraman, S.; Hsu, L.-Y.; Sachdev, V.; and Antani, S. 2022. Real-time echocardiography image analysis and quantification of cardiac indices. *Medical Image Analysis*, 102438.

Zamzmi, G.; Rajaraman, S.; Sachdev, V.; and Antani, S. 2021. Trilateral attention network for real-time cardiac region segmentation. *IEEE Access*, 9: 118205–118214.

Zhang, W.; Pang, J.; Chen, K.; and Loy, C. C. 2021. K-net: Towards unified image segmentation. *Advances in Neural Information Processing Systems*, 34: 10326–10338.

Zhang, Y.; Pang, B.; and Lu, C. 2022. Semantic Segmentation by Early Region Proxy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1258–1268.

Zhu, X.; Hu, H.; Lin, S.; and Dai, J. 2019. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9308–9316.