Deep Pyramid Local Attention Neural Network for Cardiac Structure Segmentation in Two-dimensional Echocardiography

Fei Liu , Kun Wang , Dan Liu , Xin Yang , Jie Tian

PII: S1361-8415(20)30237-1
DOI: https://doi.org/10.1016/j.media.2020.101873
Reference: MEDIMA 101873

To appear in: *Medical Image Analysis*

Received date: 28 April 2020
Revised date: 12 October 2020
Accepted date: 13 October 2020

Please cite this article as: Fei Liu , Kun Wang , Dan Liu , Xin Yang , Jie Tian , Deep Pyramid Local Attention Neural Network for Cardiac Structure Segmentation in Two-dimensional Echocardiography, *Medical Image Analysis* (2020), doi: https://doi.org/10.1016/j.media.2020.101873

HIGHLIGHTS

- A pyramid local attention module is proposed for effective feature enhancement.
- A label coherence learning mechanism is proposed to strengthen the segmentation consistency between different regions.
- PLANet achieves the joint segmentation training of unary pixel learning and pairwise correlation learning.
- PLANet efficiently performs better segmentation than other methods in 2D echocardiography.

Deep Pyramid Local Attention Neural Network for Cardiac Structure Segmentation in Two-dimensional Echocardiography

Fei Liu [a, b, *], Kun Wang [a, b, c*], Dan Liu [d, *], Xin Yang[a, b], Jie Tian[a, c, e, f, #]

[a] *CAS Key Laboratory of Molecular Imaging, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China*

[b] *Department of the Artificial Intelligence Technology, University of Chinese Academy of Sciences, Beijing, 100049, China*

[c] *Zhuhai Precision Medical Center, Zhuhai People's Hospital (affiliated with Jinan University), Zhuhai, 519000, China*

[d] *Department of Ultrasound, The Second Affiliated Hospital of Nanchang University, Nanchang, 330008, China*

[e] *Beijing Advanced Innovation Center for Big Data-Based Precision Medicine, Beihang University, Beijing, 100191, China*

[f] *Key Laboratory of Big Data-Based Precision Medicine (Beihang University)，Ministry of Industry and Information Technology, Beijing, 100191, China*

[*] *These authors contributed equally to this work and should be considered as co-first authors*

[#] *Corresponding author. E-mail address: tian@ieee.org Tel: 86-10-82628760 Fax: 86-10-62527995*

Abstract: Automatic semantic segmentation in 2D echocardiography is vital in clinical practice for assessing various cardiac functions and improving the diagnosis of cardiac diseases. However, two distinct problems have persisted in automatic segmentation in 2D echocardiography, namely the lack of an effective feature enhancement approach for contextual feature capture and lack of label coherence in category prediction for individual pixels. Therefore, in this study, we propose a deep learning model, called deep pyramid local attention neural network (PLANet), to improve the segmentation performance of automatic methods in 2D echocardiography. Specifically, we propose a pyramid local attention module to enhance features by capturing supporting information within compact and sparse neighboring contexts. We also propose a label coherence learning mechanism to promote prediction consistency for pixels and their neighbors by guiding the learning with explicit supervision signals. The proposed PLANet was extensively evaluated on the dataset of cardiac acquisitions for multi-structure ultrasound segmentation (CAMUS) and sub-EchoNet-Dynamic, which are two large-scale and public 2D echocardiography datasets. The experimental results show that PLANet performs better than traditional and deep learning-based segmentation methods on geometrical and clinical metrics. Moreover, PLANet can complete the segmentation of heart structures in 2D echocardiography in real time, indicating a potential to assist cardiologists accurately and efficiently.

Keywords: 2D echocardiography, cardiac structure segmentation, pyramid local attention, label coherence learning

1. Introduction

Cardiovascular diseases are life-threatening illnesses with high mortality rates (Chen et al., 2019). Owing to its low-cost, portability, and real-time functionality, 2D echocardiography has become an invaluable medical imaging tool in current clinical practice (Oktay et al., 2018). The segmentation of 2D echocardiographic images plays an important role in clinical routine for doctors to assess various cardiac functions, such as left ventricle volume, ejection fraction, and myocardial mass. Currently, the semi-automatic or manual annotation in cardiac ultrasound imaging is a time-consuming and operator-dependent task, which adversely affects the accuracy and efficiency of clinical diagnosis (Leclerc et al., 2019). Hence, there is a considerable demand for automatic and effective segmentation approaches to reduce the workload of cardiologists.

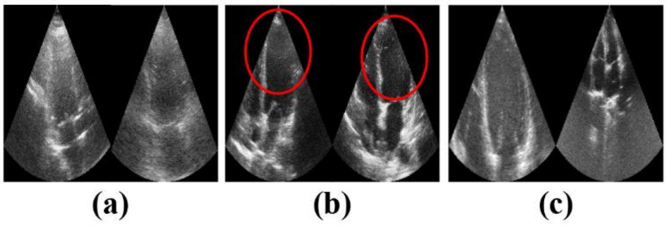

Figure 1: Examples of 2D echocardiographic images with different segmentation challenges. (a) Low signal-to-noise ratios and speckles. (b) Low image contrast between the ventricle and myocardium. (c) Large shape variability of cardiac structures.

However, multiple intrinsic limitations make the automatic segmentation of echocardiographic images an open and difficult task. The low signal-to-noise ratio and speckles hinder the robustness of segmentation methods (Figure 1a). Additionally, poor image contrast between the blood pool and myocardium makes it difficult to determine the contour of the ventricles. Accurate segmentation of the ventricles, especially the left ventricle, is crucial for several quantitative measures (Leclerc et al., 2019) (Figure 1b). Moreover, the significant shape variability of cardiac structures in 2D echocardiographic images is a crucial test for the generalization of automatic segmentation approaches (Figure 1c).

Owing to the importance and challenges of segmentation in echocardiography mentioned above, the automation of this task has been an important research topic in the past decades. For traditional segmentation methods based on energy optimization, some prior information forms have been found useful in regularization for accurate segmentation. They include the shape (Davatzikos et al., 2003; Pedrosa et al., 2017), tissue appearance (van Stralen et al., 2014; Wang and Smedby, 2014), atlas (Oktay et al., 2014), and cardiac motion (Smistad and Lindseth, 2014). Machine learning algorithms based on feature engineering have also been proposed to segment cardiac structures in echocardiography. Leclerc et al. (2017) applied structured random forest by incorporating different contextual information of multiple scales to accurately segment the myocardium and left ventricle. Based on sparse representation and dictionary learning, Huang et al. (2014) exploited the spatiotemporal coherence of echocardiographic sequences to constrain the estimation of the cardiac contour. These traditional methods are usually based on handcrafted features or raw echo intensity, whose expressiveness hinders further accuracy improvements.

Recently, with the success of deep learning in numerous medical imaging fields, several researchers have exploited the application of deep learning-based methods on echocardiography segmentation. Inspired by the effectiveness of prior information in traditional algorithms, priors such as shape (Oktay et al., 2018) and atlas (Dong et al., 2020) have been incorporated into deep learning frameworks to provide anatomical structure knowledge, and they achieved encouraging improvements compared to models without priors. Methods combining deep learning with deformable models have also been developed for two-step (Carneiro et al., 2012; Veni et al., 2018) and one-step (Nascimento and Carneiro, 2019) heart structures segmentation. Motion patterns in cardiac ultrasound sequences have been explored by particle filtering (Carneiro and Nascimento, 2013) and optical flow (Jafari et al., 2018) to maintain the temporal coherence for consistent and accurate segmentation on different timestamps. Moreover, several researchers have focused on utilizing unlabeled (Carneiro and Nascimento, 2012; Yu et al., 2017) and multi-domain (Chen et al., 2016) data in the training phase to reduce the requirement of deep learning-based models for large medical imaging datasets.

Despite these efforts, two major problems remain unsolved in the field of segmentation in echocardiography. First, low-contrast between myocardium tissues and edge dropout are commonplace in 2D echocardiography (Figure 1). Thus, a specific model design is required to enhance the features of low-contrast regions based on neighboring contexts while reducing the negative impact of noises. Second, the current deep learning-based segmentation methods typically predict the category for each pixel independently. A prediction for one pixel is made without explicitly considering other prediction results of neighboring pixels. Thus, they lack the learning mechanism for label coherence between different locations in a 2D echocardiographic image, which is likely to be suboptimal and reduces segmentation quality (Chen et al., 2018).

In the current deep learning framework, the attention mechanism is a common tool to integrate contexts directly for feature enhancement, and its efficacy has been widely proven in several research fields (Vaswani et al., 2017; Xia et al., 2019). Most applications of the attention mechanism, especially in the field of computer vision, are global attention, which means that one point in the input has access to all the other positions of the input with the estimated combination weights. Global attention has been applied widely in many studies for feature enhancement (Bian et al.; Lei et al., 2020a; Wang et al., 2018; Xu et al., 2020) and information fusion of multi-source data (Wang et al., 2020; Wang et al., 2019a; Wang et al., 2019b). It is also important for some deep learning-based algorithms to utilize global attention for improving the medical interpretability of network features (Araújo et al., 2020; Guo and Yuan, 2020; Stolte and Fang, 2020; Wei et al., 2019) or localizing lesions in a weakly supervised manner (Lei et al., 2020b; Maicas et al., 2019; Pesce et al., 2019). In contrast, “local attention,” which restricts the context in compact and reasonable scales, has not been sufficiently studied in image processing methods. Brain science studies (Crewther et al., 2007; Heinze and Münte, 1993) have proven that local attention is crucial and is processed in parallel with global attention via independent neural pathways at the early stages of cortical processing. Local attention has been applied successfully in several studies on machine translation (Luong et al., 2015), speech recognition (Mirsamadi et al., 2017), and image generation (Gregor et al., 2015), achieving significant improvement compared to global attention-based methods. However, in the field of semantic segmentation, the successful application of local attention is still rare. Currently, in the field of medical image analysis, local attention mainly works with graph neural networks for collecting contextual cues from neighboring regions. For example, Khosravan et al. (2019) developed a collaborative CAD system to analyze the visual search pattern of radiologists using local attention for data sparsification. To segment vessel structures, Shi et al. (2019) applied local attention to exploit the graphical structure of vessel shape and improved vessel segmentation accuracy. We argue that a properly designed local attention aids the segmentation in contour regions, where the association of features in the neighborhood is much stronger than features far away from the segmentation positions. Moreover, owing to the low signal-to-noise characteristics of echocardiography, global attention is prone to being overwhelmed by the cumulative imaging noise, which can be avoided in local attention by excluding unrelated contexts explicitly.

Since proposed by Long et al. (2015), fully convolutional networks (FCNs) have been the mainstream of deep learning-based segmentation methods. The aim of an FCN segmentation is to assign each pixel with a category independently (Liu et al., 2018), while overlapping the receptive fields of pixels significantly to improve the efficiency of feedforward and backpropagation computations (Shelhamer et al., 2017). As there are strong semantic correlations among pixels in an image, it is important to enhance label agreement for similar pixels and label disagreement for dissimilar pixels in semantic segmentation models. Conditional random field (CRF) and Markov random field (MRF) have been the most common tools for modeling the joint distribution of labels for pixels. For example, Chen et al. (2018) proposed introducing CRF in the post-processing step to refine the segmentation results of FCN. Meanwhile, Zheng et al. (2015) and Liu et al. (2018) formulated the CRF and MRF into FCN, respectively, achieving end-to-end network training and encouraging improvement. However, as a consequence of this incorporation, the network training was very slow and complex.

In this study, we developed a novel method to enhance label consistency for different pixel pairs. To achieve this, we first analyzed the challenges that inhibit the segmentation performance of deep learning-based methods in 2D echocardiography. Then, we developed our proposed method called deep pyramid local attention neural network (PLANet) to solve them. Specifically, we parse the connection pattern of one pixel with its neighbors in label maps as supervision signals to guide the FCN in learning pairwise label correlation explicitly. Meanwhile, the segmentation result is updated by the neighboring prediction with the learned label correlation as weights.

Our contributions in this work are summarized as follows:

1. We propose a pyramid local attention (PLA) module to exploit context locality for feature enhancement in contour regions, capturing supporting features within compact neighboring contexts under low imaging contrast and high noise.

2. We propose a novel label coherence learning (LCL) mechanism to explicitly learn the prediction consistency of pixels and their neighbors, which achieves the joint training of unary pixel learning and pairwise correlation learning, improving the segmentation quality in 2D echocardiography.

3. We evaluate the segmentation performance of the proposed PLANet on the cardiac acquisition for multi-structure ultrasound segmentation (CAMUS) (Leclerc et al., 2019) and sub-EchoNet-Dynamic (Ouyang et al., 2020) datasets, which are two very large-scale and public 2D echocardiography datasets (Chen et al., 2019). Compared to state-of-the-art methods, PLANet achieved better segmentation on geometrical and clinical evaluation metrics. Moreover, PLANet is efficient and completes the cardiac structure segmentation of 2D echocardiography images in real time, indicating its potential as a tool for assisting cardiologists in clinical practice.

The rest of this paper is organized as follows: Section 2 illustrates the details of the backbone network of PLANet, proposed PLA module, LCL mechanism, and training methods. Section 3 presents our experimental settings and results, and we discuss the properties of the proposed PLANet in Section 4. The conclusions are presented in Section 5.

2. Methodology

In this work, given a 2D echocardiographic image as input, the proposed PLANet estimates the segmentation results for the left ventricle endocardium and myocardium, where the size of the segmentation output is the same as the input. As illustrated in Figure 2, the backbone network extracts deep semantic features for the input. A primary category estimation is then assigned for each pixel separately, using the PLA module for feature enhancement within the compact and sparse neighboring contexts. Finally, in the proposed LCL mechanism, the primary segmentation result of one pixel is propagated by the coherence with the neighboring pixels, where the coherences are estimated in advance by the network. PLANet is described in detail in the following subsections.

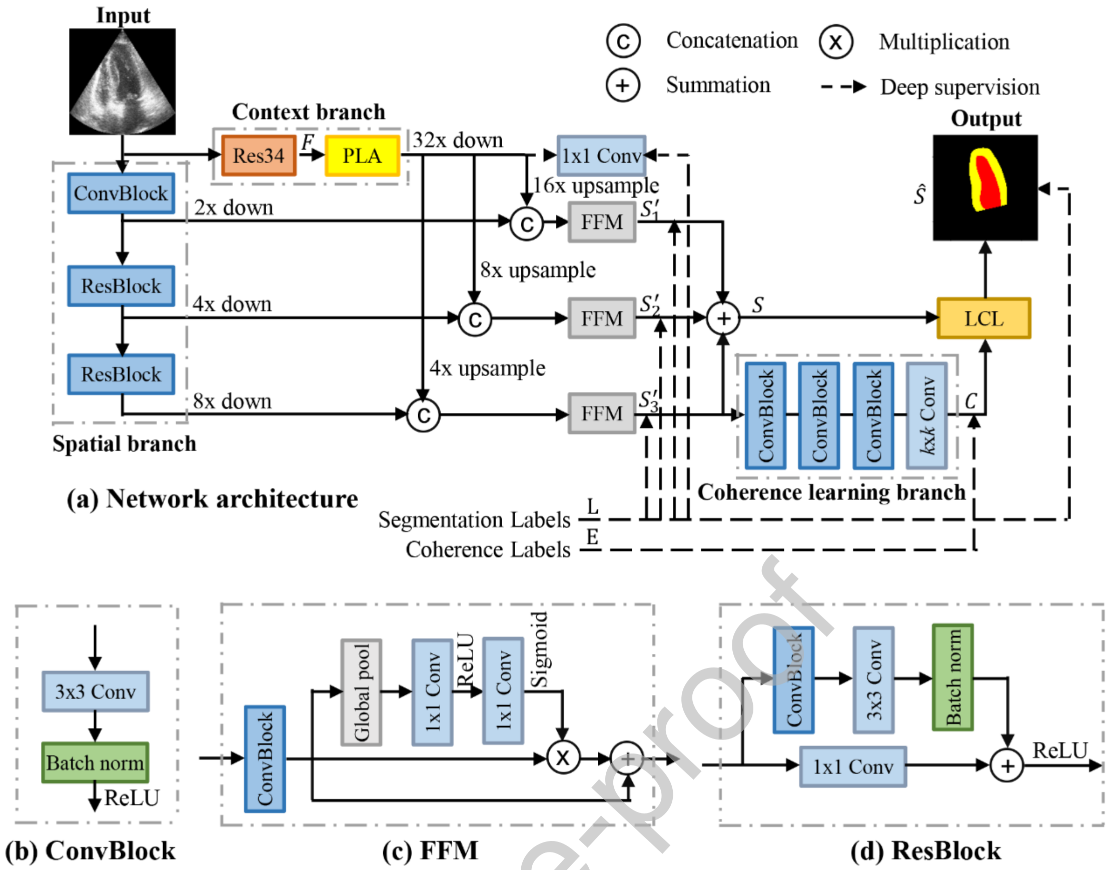

Figure 2: Overview of the proposed PLANet. (a) Network architecture of PLANet. The input image is passed into context and spatial branches to extract different types of features. Features in the context branch are enhanced by the PLA module within neighboring contexts. Primary segmentations are predicted by fusing the features from two branches. The coherence learning branch and LCL module further update the segmentations, while considering the label coherence between different regions. (b) Convolutional block. (c) Feature fusion module. (d) Residual convolutional block.

2.1 Backbone network

It is important for semantic segmentation models to preserve enough spatial information while offering a sufficiently large receptive field. Spatial information ensures detail sharpness in the segmentation, and a large receptive field ensures the correctness of the assigned category. In this work, we adopted the design philosophy of BiSeNet (Yu et al., 2018) for building the backbone network, and made necessary modifications to fit the characteristics of echocardiographic images.

BiSeNet comprises a spatial branch and context branch, which extract the necessary spatial details and abundant semantic features, respectively. ResNet34 (He et al., 2016), pretrained on ImageNet, was adopted in the context branch as the feature extractor. Cardiac structures, such as the left ventricle, are larger in 2D echocardiography than the objects in natural images of the same size. Hence, we doubled the dilation of the convolutional layers in the last residual block of ResNet34 to expand the receptive field of the context branch. In the spatial branch, two residual blocks used in the last two layers reduce the vanishing of training gradients and loss of spatial details (Peng et al., 2017). The output feature map size of the spatial branch is 1/8 of the original image.

To fully capture the multi-scale spatial details, we extracted the network features with different resolutions from the output of three layers in the spatial branch. The semantic contextual features from the context branch were then merged with the three spatial features by a dedicated feature fusion module (FFM), and three temporary segmentation results S_1', S_2' and S_3', respectively, were predicted. Channel attention (Hu et al., 2019) was applied in the FFM to model the interdependencies between the channels of merged features. A primary segmentation map S was learned by $S = (S_1' + S_2' + S_3')/3$.

2.2 Pyramid local attention

Empirically, the feature coherence of one pixel and its neighboring region is considerably stronger compared to the distant regions. The widely used global attention mechanism compares the feature affinity of one point with all other positions indiscriminately, which disperses the distribution of attention and leads to coarser-grained features (Li et al., 2018). Moreover, as the signal-to-noise of echocardiography is lower than that of other medical imaging methods, such as CT or MRI (Chen et al., 2019), meaningful information might be weakened by the stacked noise. Therefore, we proposed a PLA module to enhance features within restricted and reasonable contexts to overcome the shortcomings of the traditional global attention.

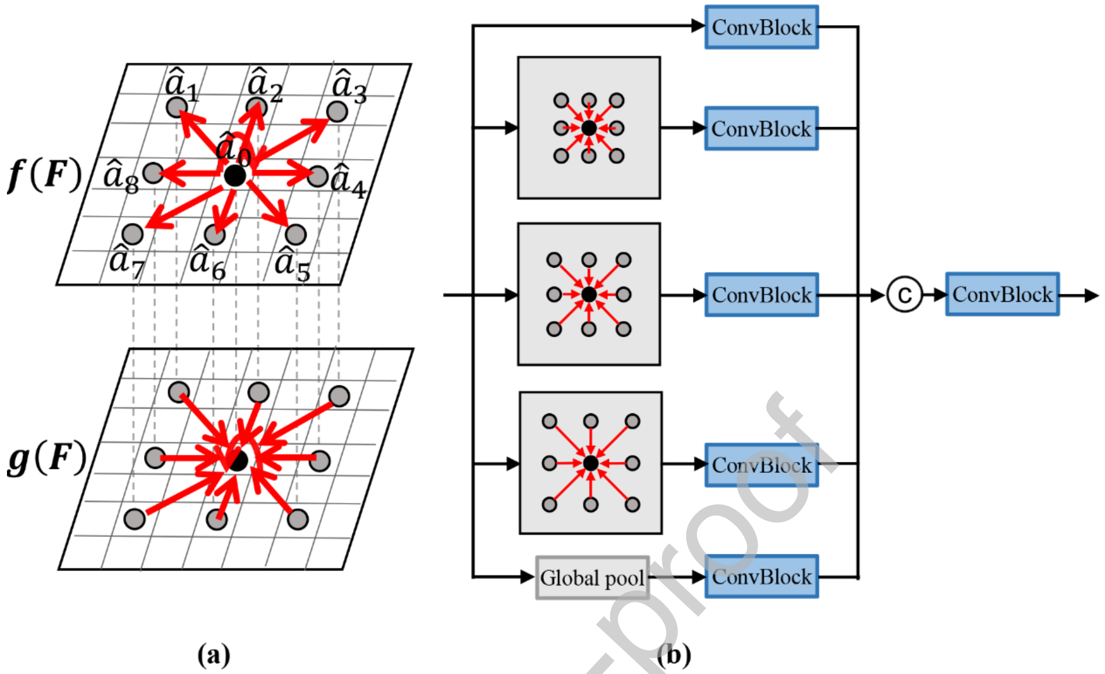

Figure 3: Schema of the PLA module. (a) In local attention, feature affinities for the central positions are estimated within compact and sparse neighboring contexts. Then, the features in the central positions are enhanced by collecting contextual features with attention weights. (b) We combine multiple local attention modules with different sparsities as PLA module. We take sampling position number $d = 3$ and sampling stride $s = 2$ as examples in this figure.

The semantic features F are extracted from the context branch of the backbone network, where $F \in \mathbb{R}^{c \times h \times w}$; c, h, and w represent the number of channels, height, and width of feature F, respectively. We now describe the local attention (LA) applied on feature F (Figure 3a). To save computation in the attention mechanism, embedding functions f and g project feature F into two new feature spaces to reduce the channel dimension of F. Here, we implemented functions f and g as convolutional layers with 1×1 kernels. The feature affinity for a pair of feature positions p_1 and p_2 is defined as follows:

$$a(p_1, p_2) = \langle f(F_{p_1}), f(F_{p_1}) \rangle \quad (1)$$

In (1), $\langle \cdot \rangle$ represents the dot product. The sampling of feature position p_2 is constrained within the region of $\Phi(p_1; d, s)$, where p_1 is the center of the sampling region, d is the number of sampling positions along each axis, and s is the sampling stride, which is the minimum distance between two available sampling positions. Thus, we defined a compact and sparse neighboring region for the local attention mechanism in this study. The feature affinities a within one sampling region Φ are then normalized as attention weights $\hat{a}$ by the softmax function:

$$\hat{a}(p_1, p_2) = \frac{exp(a(p_1, p_2))}{\sum_{p \in \Phi(p_1; d, s)} exp(a(p_1, p))} \quad (2)$$

With the introduction of normalization, the more related features are strengthened while unrelated features are weakened to contribute less when incorporating the contextual information for position p_1. Further, the feature in position p_1 is enhanced by the contextual features in Φ with the corresponding attention weights $\hat{a}$:

$$\hat{F}_{p_1} = \sum_{p \in \Phi(p_1; d, s)} \hat{a}\,(p_1, p) g(F_p) \quad (3)$$

Inspired by Vaswani et al. (2017), we adopted multi-head local attention in our study; consequently, the final enhanced feature of local attention $LA(F)$ is the concatenation of results from multiple independent local attentions $LA_i(F)$:

$$LA(F) = Concat(LA_1, LA_2, \cdots, LA_h) \quad (4)$$

where we used four heads of local attention in our experiments.

Owing to the significant size variability of cardiac structures within the population (Leclerc et al., 2019), one local attention can only cover a fraction of possible cases. Therefore, we processed the semantic feature F in a pyramid of local attentions with different sampling strides $s = (2,4,8)$. In this way, we maintained the benefits of local attention while being robust for the segmentation of multi-sized heart structures. As illustrated in Figure 3b, in parallel with the PLA, we applied global average pooling for capturing image-level features and a plain convolution layer for maintaining the finest-grained features (Chen et al., 2017).

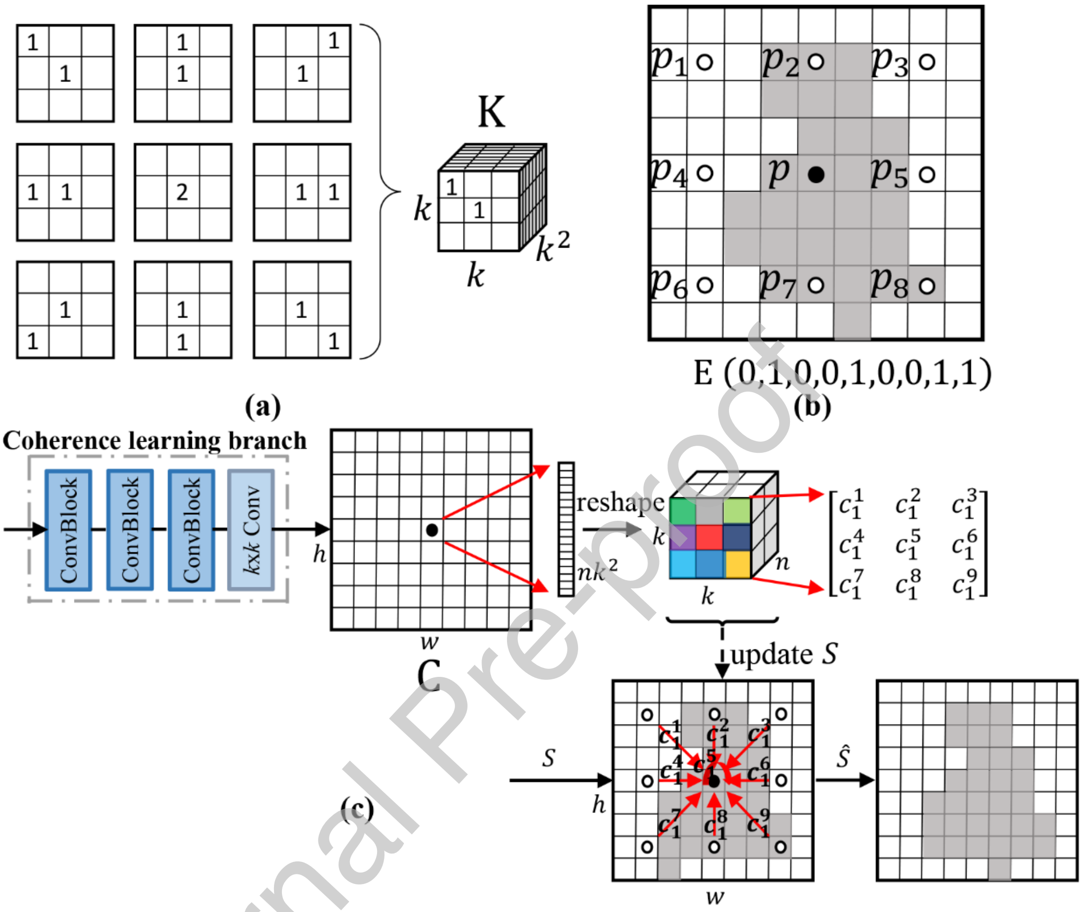

Figure 4: LCL mechanism. (a) Design of the convolution kernel for label coherence parsing. (b) Parsed label coherence E for position p within a neighboring region. (c) Design of LCL module for updating segmentations with estimated coherences C. We take kernel size $k = 3$, dilation $m = 3$, and category number $n = 2$ as examples in this figure.

2.3 Label coherence learning mechanism

Most semantic segmentation methods predict the category for each pixel independently without an explicit mechanism to learn the label coherence for one pixel and its neighbors (Taghanaki et al., 2019). This approach is suboptimal and prone to poor contour delineation and scattered spurious regions in the segmentation results (Zheng et al., 2015). Therefore, in this present study, we proposed a novel LCL mechanism to solve this problem.

The aim of LCL is to consider the semantic consistency within neighbors when assigning a category for each position. To lead the network to explicitly learn the consistency between pixels, we designed specific supervision signals based on the segmentation label maps. We first converted the segmentation label map $\mathrm{L} \in \mathbb{R}^{1\times h\times w}$ with n segmentation categories into a tensor $\mathrm{V} \in \mathbb{R}^{n\times h\times w}$, making the label of each position in V a one-hot vector. A specific convolution layer with manually determined kernel $\mathrm{K} \in \mathbb{R}^{k^2\times 1\times k\times k}$ was applied to parse the label coherence for pixel p_1 in the neighboring context $\Psi(p_1; k, m)$, where k represents the number of sampling positions along each axis, whose coherence requires consideration with the center of neighboring context region Ψ, and the sampling stride m represents the minimum distance between sampling positions (Figure 4a). The determination of kernel K was as follows:

$$\begin{aligned} &\mathrm{K}(i, 1, \lfloor i/k \rfloor, i \bmod k) = 1 \quad i \in (1,2,\cdots,k^2) \\ &\mathrm{K}(:,:,\lfloor k/2 \rfloor, \lfloor k/2 \rfloor) = \mathrm{K}(:,:,\lfloor k/2 \rfloor, \lfloor k/2 \rfloor) + 1 \end{aligned} \tag{5}$$

where $\lfloor\cdot\rfloor$ indicates the rounding down and mod represents the modulus operation. We convolved each channel of the transformed segmentation map V with kernel K separately, and the factor m in $\Psi(p; k, m)$ was used as the dilation of the convolution. Each value in the convolution result $\mathrm{E} \in \mathbb{R}^{n\times k^2\times h\times w}$ was binarized to indicate whether the sampling position possesses the same foreground category with the center position of a neighboring region: non-zero represents the same category, and zero represents a different category or both are background category (Figure 4b).

Further, we added an extra coherence learning branch for the prediction of label coherence, which is in parallel with the prediction of primary segmentation S. The coherence learning branch consisted of four convolutional layers, with kernel of 3 and dilation of 2 for the first three layers to offer large sufficient receptive fields. For the last convolution layer, the number of output channel was set as nk^2, the kernel size as k, and dilation as m, keeping the same parameters of the segmentation label maps parsing. The output vector for each position p concatenated the estimation of all label coherences $C_l(p, p_i)$ with the corresponding neighboring positions p_i under all categories l. We supervised the prediction of the coherence learning branch by the coherence labels E for learning expected results.

With the estimation of label coherence for pairs of positions, we propagated the segmentation $S_l(p)$ of one position p under category l within its neighboring context $\Psi(p; k, m)$ (Figure 4c). Specifically, the primary segmentations of position p were updated by the segmentation combination of the sampling positions in $\Psi(p; k, m)$, with the weights as the corresponding coherence strength $C_l(p, p_i)$ predicted by the coherence learning branch:

$$\hat{S}_l(p) = \frac{1}{N} \sum\nolimits_{p_i \in \Psi(p;k,m)} S_l\,(p_i) C_l(p, p_i) \tag{6}$$

where the normalization factor $N = \sum_{p_i \in \Psi(p;k,m)} C_l(p; p_i)$, and $\hat{S}_l(p)$ is the updated segmentation for position p under category l. The updated segmentation $\hat{S}_l$ was supervised in the training phase along with the primary segmentation S_l. In this way, we connected the segmentation predictions across different positions with an explicit supervision signal to guide the efficient label coherence learning.

2.4 Training of PLANet

In this work, we are required to learn two kinds of prediction tasks: semantic segmentation and label coherence learning. For the segmentation task, we adopted deep supervision (AI-Barazanchi et al., 2016) to train the proposed PLANet. A principle loss function L_{SP} was used to supervise the learning of updated segmentation $\hat{S}_l$, and multiple auxiliary loss functions L_{SA} guided the learning of three temporary segmentations S'_1, S'_2, and S'_3. Both L_{SP} and L_{SA} are cross-entropy loss functions:

$$L_{SP} = L_{SA} = -\frac{1}{N}\sum_{i=1}^{N} L(p_i) log\left(\frac{exp\big(S(p_i)\big)}{\sum_{j=1}^{N} exp\left(S\big(p_j\big)\right)}\right) \tag{7}$$

where N is the total number of pixels in segmentation map S, and L is the ground truth of the segmentation.

For the label coherence learning task, we adopted the binary cross-entropy loss function L_C to supervise the coherence estimation C with label E parsed from the segmentation label maps, as described in Section 2.3.

The total loss function $Loss$ for PLANet is the joint loss of L_{SP}, three L_{SA}s, and L_C:

$$Loss = L_{SP} + \alpha \cdot (L_{SA1} + L_{SA2} + L_{SA3})/3 + \beta \cdot L_C \tag{8}$$

where α and β balance the weights of the three loss items. α and β in our experiments are equal to 0.5 empirically. As the segmentation label maps in L_{SP} and L_{SA} are the same size as the input image, and the label coherences E in L_C are created by convolution run on GPU, the labels used in our study can be generated conveniently and efficiently.

We trained the proposed PLANet using a mini-batch stochastic gradient descent (SGD) solver, with momentum 0.9, weight decay 10^{-4}, and batch size 24. We trained PLANet with 500 epochs. Before the network training, we warmed up the model by freezing the parameters of ResNet34 and pre-training other parameters by 20 epochs with a learning rate of 10^{-4}. We resized the 2D echocardiographic image to 512×512 before inputting it to PLANet. The learning rate was decayed in a cosine annealing schedule from 10^{-1} to 10^{-4}. As the memory-consumption in the training phase was greater than the total memory of our hardware, we adopted a training trick of cumulating backward computing gradients twice before updating the network weights and clearing the gradients. We set the decay of batch normalization parameters as 0.9997, as suggested in DeepLabv3 (Chen et al., 2017). We further applied data augmentation by random horizontal flip, random rotation from -20° to 20°, random scaling of the input image from 0.7 to 1.3, and random translation from -0.3 to 0.3.

The experiments were conducted on one computer equipped with an Intel I9-9900K 3.6 GHz CPU, two NVIDIA TITAN V GPUs, and 64 GB RAM. We implemented the proposed PLANet in Python (v. 3.6) with the PyTorch (v. 1.3.0) toolkit (Paszke et al., 2017). Specifically, the operations of neighboring context sampling in PLA and LCL modules are achieved by the Fold and Unfold functions in PyTorch. Our PLANet with GPU implementation runs at 61fps in the inference of segmentation using a single thread.

3. Experiment settings and results

3.1 Data

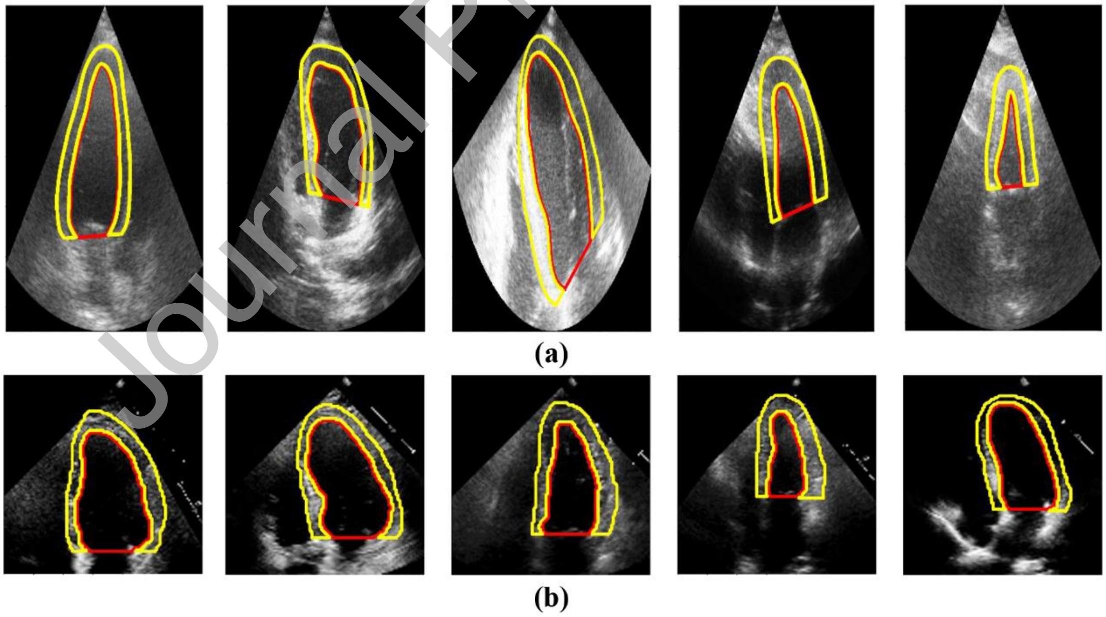

Figure 5: Annotation examples of the CAMUS dataset (a) and sub-EchoNet-Dynamic dataset (b). The red and yellow lines represent the contours of left ventricle and myocardium, respectively.

CAMUS dataset: We evaluated the proposed PLANet on the CAMUS dataset (Figure 5a), which is an open large-scale dataset in 2D echocardiography (Leclerc et al., 2019). CAMUS maintains a wide heterogeneity of image quality and pathological cases to preserve the clinical realism. Five hundred patients were enrolled in the CAMUS dataset. The organizers categorized the patients based on image quality into *Good* (175), *Medium* (230), and *Poor* (95). Echocardiographic images were acquired from GE Vivid E95 ultrasound scanners with GE M5S probe. The apical four-chamber and two-chamber view sequences were exported for each patient and annotated manually at end diastole (ED) and end systole (ES). Thus, there are 2000 echocardiographic images in CAMUS dataset. Cardiac structures with manual annotations are left ventricle endocardium (LV_{Endo}) and myocardium (LV_{Epi}). Three cardiologists participated in the manual annotation for the 2D echocardiographic images in the CAMUS dataset, and they strictly followed the same segmentation protocol. The annotations of 50 patients in the CAMUS dataset are not accessible, and their segmentations need to be evaluated online. Please refer to Leclerc et al. (2019) for more detailed information about the CAMUS dataset.

sub-EchoNet-Dynamic dataset: We built a new cardiac image segmentation dataset by randomly sampling 2500 echocardiogram videos from EchoNet-Dynamic (Ouyang et al., 2020), which is a large-scale public dataset for assessing cardiac functions. We named the new dataset as sub-EchoNet-Dynamic (Figure 5b). The distributions of enrolled population between CAMUS and sub-EchoNet-Dynamic are different (Figure S1). The ratio of population in pathological risk in CAMUS is higher than that in sub-EchoNet-Dynamic (Table S1). The ultrasound machines used to acquire the data in sub-EchoNet-Dynamic include Acuson SC2000, Epiq 5G, Epiq 7C, iE33, and Sonos (Ouyang et al., 2020). In sub-EchoNet-Dynamic, one apical four-chamber sequence was exported for each patient and saved as an AVI file for privacy protection. LV_{Endo} and LV_{Epi} at ED and ES were annotated manually by experienced cardiologists and checked carefully by one neutral observer. We extracted the frames at ED and ES for the experiment in this work. The dataset was randomly divided into training, validation, and testing datasets with 1600, 400, and 500 patients, respectively. Please refer to Ouyang et al. (2020) for a more detailed design protocol of the EchoNet-Dynamic dataset. The main characteristics of the CAMUS and sub-EchoNet-Dynamic datasets are given in the Supplementary Materials.

3.2 Evaluation metrics

We used geometrical and clinical metrics to quantitatively evaluate the segmentation performance of the proposed PLANet, which are all well reported in the literature (Chen et al., 2019; Leclerc et al., 2019). Based on the suggestion of Leclerc et al. (2019), cardiac images with poor quality were excluded in the performance evaluation as they are clinically useless.

The geometrical metrics contain the Dice (D), mean absolute distance (d_m), and 2D Hausdorff distance (d_H). The Dice metric D measures the overlap ratio between the automatic segmentation S_m by methods and ground-truth of manual annotations S_g by cardiologists. The measurements were rated from 0 (worst) to 1 (best).

$$D\left(S_g, S_m\right) = 2\left|S_g \cap S_m\right| / \left(\left|S_g\right| + \left|S_m\right|\right) \tag{9}$$

d_m measures the mean surface distance between automatic segmentations and annotations, and d_H corresponds to the maximum distance between the contours of automatic segmentations and manual annotations. When the values of d_m and d_H are lower, the automatic segmentation is better.

$$\begin{gathered} d_m\left(S_g, S_m\right) = \max\left(\frac{1}{\left|S_g\right|}\sum_{a \in S_g} \min_{b \in S_m} \|a - b\|, \frac{1}{\left|S_m\right|}\sum_{b \in S_m} \min_{a \in S_g} \|b - a\|\right) \\ d_H\left(S_g, S_m\right) = \max\left(\max_{a \in S_g} \min_{b \in S_m} \|a - b\|, \max_{b \in S_m} \min_{a \in S_g} \|b - a\|\right) \end{gathered} \tag{10}$$

The geometrical metrics for echocardiographic images in ED and ES were measured separately.

Three clinical indices are frequently used by cardiologists in clinical practice: volume of the left ventricle in both ED and ES ($\mathrm{LV_{EDV}}$ and $\mathrm{LV_{ESV}}$) and ejection fraction ($\mathrm{LV_{EF}}$). $\mathrm{LV_{EDV}}$ and $\mathrm{LV_{ESV}}$ are computed with the Simpson's rule.

$$\mathrm{LV_{EF}} = (\mathrm{LV_{EDV}} - \mathrm{LV_{ESV}})/\mathrm{LV_{EDV}} \tag{11}$$

In this study, the clinical metrics were the Pearson correlation coefficient (*corr*), mean bias (*bias*), and standard deviations (σ) of the three clinical indices, respectively. We computed the clinical metrics for the patients with accessible annotations.

3.3 Ablation experiments

We investigated the contributions of each component to the segmentation performance of the proposed PLANet. We also explored the impact of some important hyper-parameters on the behavior of PLANet. In the ablation experiments, the patients with annotations in CAMUS dataset were randomly divided into training (350) and evaluation (100) datasets. All the methods in this subsection were trained under the same settings described in Section 2.4. We determined the model with the best performance by voting on all metrics. The full details of the ablation experiments on the two datasets are presented in the Supplementary Materials.

Ablation for PLA module: We investigated the influence of the sampling number d and sampling stride s on the PLA. The sampling number d defines the size of the neighboring context of PLA while the sampling stride s controls the sparsities of the local attentions. The network structure contained only backbone network and PLA in this experiment. First, we fixed $d = 3$ to compare the effectiveness of PLA with different sampling strides s. The segmentation performances under various values of s are listed in Table 1. "Too sparse" or "too dense" is harmful for the PLA to collect neighboring contexts efficiently, and $s = (2,4,8)$ is the best combination in this work. Then, we set $s = (2,4,8)$ and tested the optimum sampling number d. As shown in Table 2, restricting the size of a neighboring context improves the segmentation performance. The improvement reaches its peak when $d = 5$, after which it worsens slightly when $d = 3$, indicating the insufficiency of contextual information in this case.

Table 1: Segmentation performance for LV$_{\text{Endo}}$ and LV$_{\text{Epi}}$ under different sampling strides of PLA.

Sampling strides*		ED			ES		
		D	d_m(mm)	d_H(mm)	D	d_m(mm)	d_H(mm)
LV$_{\text{Endo}}$	(2,4,6)	0.949±0.020	1.31±0.48	4.37±1.72	0.921±0.042	1.49±0.68	4.88±1.86
	(2,4,8)	**0.950±0.018**	**1.29±0.44**	4.42±1.76	**0.926±0.033**	**1.39±0.57**	**4.49±1.53**
	(2,6,8)	0.948±0.018	1.33±0.47	**4.21±1.78**	0.923±0.038	1.47±0.63	4.76±1.82
	(2,4,10)	0.947±0.021	1.31±0.53	4.55±2.11	0.917±0.043	1.58±0.75	4.91±1.76
LV$_{\text{Epi}}$	(2,4,6)	0.955±0.014	**1.65±0.57**	5.22±2.25	0.943±0.023	1.89±0.77	5.13±1.84
	(2,4,8)	**0.956±0.015**	1.67±0.59	5.23±2.49	**0.945±0.021**	**1.87±0.77**	**5.07±1.66**
	(2,6,8)	0.955±0.015	1.71±0.63	**5.18±2.50**	0.942±0.022	1.96±0.73	5.24±1.71
	(2,4,10)	0.952±0.016	1.75±0.57	5.35±2.79	0.941±0.023	1.88±0.75	5.12±1.78

* ED: end diastole; ES: end systole; D: Dice; d_m: mean absolute distance; d_H: Hausdorff distance; LV$_{\text{Endo}}$: endocardial contour of the left ventricle; LV$_{\text{Epi}}$: epicardial contour of the left ventricle.

Table 2: Segmentation performance for LV$_{\text{Endo}}$ and LV$_{\text{Epi}}$ under different sampling sizes of PLA.

Sampling sizes		ED			ES		
		D	d_m(mm)	d_H(mm)	D	d_m(mm)	d_H(mm)
LV$_{\text{Endo}}$	3	0.950±0.018	1.29±0.44	4.42±1.76	0.926±0.033	1.39±0.57	4.49±1.53
	5	**0.953±0.020**	**1.25±0.50**	**4.39±1.86**	**0.927±0.035**	**1.37±0.58**	**4.42±1.74**

	7	0.951±0.017	1.27±0.41	4.48±1.69	0.925±0.033	1.42±0.58	4.72±1.84
	9	0.948±0.019	1.31±0.47	4.46±1.74	0.921±0.037	1.48±0.61	4.71±1.81
LV_{Epi}	3	0.956±0.015	1.67±0.59	5.23±2.49	0.945±0.021	**1.87±0.77**	5.07±1.66
	5	**0.958±0.014**	1.66±0.62	**4.99±2.16**	**0.946±0.022**	1.90±0.78	5.15±1.92
	7	0.956±0.014	**1.59±0.57**	5.15±2.16	0.944±0.021	1.93±0.71	**5.07±1.63**
	9	0.955±0.015	1.72±0.59	5.27±2.46	0.942±0.024	1.97±0.82	5.32±1.75

Ablation for LCL mechanism: It is important to properly adjust the scale of context and sparsity for the LCL mechanism. The coherence labels are full of homogeneity when the neighboring regions are too narrow and dense. Likewise, a setting that is too large and sparse makes the learning task extremely easy, as the dissimilarities of most connected pixel pairs are significantly clear. In this experiment, we tested the influence of different context scales k and sparsity values m on the segmentation performance of the network. We used the model structure composed of the backbone network and the LCL mechanism. Similar to the ablation for PLA, we fixed the context scale to k= 5 and evaluated the performance under the sparsity $m = (12,14,16,18)$. The performances are summarized in Table 3. The proper sparsity $m = 16$ was determined empirically. Then, we investigated the effect of the context scale when $k = (3,5,7,9)$ and $m = 16$. The results shown in Table 4 indicate that the best context scale is $k = 7$ under our experimental settings.

Table 3: Segmentation performance for LV_{Endo} and LV_{Epi} under different sampling strides of LCL mechanism.

Sampling strides		ED			ES		
		D	d_m(mm)	d_H(mm)	D	d_m(mm)	d_H(mm)
LV_{Endo}	12	0.944±0.025	1.42±0.61	4.65±2.03	0.913±0.042	1.60±0.61	5.05±2.12
	14	0.950±0.022	1.31±0.55	4.39±1.87	0.918±0.046	1.52±0.65	4.88±2.21

	16	**0.951±0.017**	**1.26±0.41**	4.37±1.68	**0.920±0.031**	**1.37±0.52**	**4.56±1.68**
	18	0.951±0.019	1.29±0.48	**4.35±1.71**	0.917±0.037	1.46±0.61	4.72±2.00
LV_{Epi}	12	0.951±0.015	1.60±0.67	5.13±2.77	0.942±0.022	**1.79±0.72**	5.26±2.05
	14	0.953±0.015	**1.57±0.62**	5.05±2.62	0.943±0.023	1.86±0.82	5.28±2.11
	16	**0.955±0.014**	1.59±0.59	**4.92±2.24**	**0.945±0.023**	1.84±0.74	**5.21±1.84**
	18	0.953±0.013	1.61±0.58	5.03±2.15	0.944±.0.025	1.82±0.76	5.30±1.87

Table 4: Segmentation performance for LV_{Endo} and LV_{Epi} under different sampling sizes of LCL mechanism.

Sampling sizes		ED			ES		
		D	d_m(mm)	d_H(mm)	D	d_m(mm)	d_H(mm)
LV_{Endo}	3	0.950±0.020	1.31±0.48	4.39±1.59	0.919±0.041	1.49±0.64	4.87±2.11
	5	0.951±0.017	**1.26±0.41**	4.37±1.68	0.920±0.031	**1.37±0.52**	4.56±1.68
	7	**0.952±0.018**	1.26±0.45	**4.34±1.65**	**0.924±0.037**	1.47±0.61	**4.51±1.77**
	9	0.950±0.020	1.28±0.49	4.38±1.61	0.918±0.038	1.53±0.64	4.83±1.94
LV_{Epi}	3	0.954±0.013	1.65±0.55	4.95±2.03	0.943±0.023	1.85±0.75	5.25±1.94
	5	0.955±0.014	**1.59±0.59**	**4.92±2.24**	0.945±0.023	1.84±0.74	5.21±1.84
	7	**0.957±0.013**	1.63±0.60	5.05±2.02	**0.947±0.024**	**1.80±0.80**	**5.13±1.82**
	9	0.955±0.015	1.67±0.63	5.15±2.12	0.944±0.025	1.94±0.78	5.43±2.04

Ablation for the contribution of each component: We tested the performance improvement when adding components step-by-step. Different components have different levels of performance contributions to PLANet. We also evaluated the performances when the PLA was replaced by atrous spatial pyramid pooling (ASPP) or multi-head global attention (GA). In this way, we compared the PLA module against the plain convolution and global attention. Based on the results of the ablation experiments for PLA and LCL modules, we set $d = 5, s = (2,4,8)$ for PLA and $k = 7, m = 16$ for LCL in this experiment. The segmentation accuracy of all the models in this experiment is listed in Table 5. A significant improvement can be observed with the addition of the PLA and LCL on the backbone network gradually. The performance of the global attention was similar to those of ASPP in our experiments, whereas both achieved context collection differently. The PLA worked better than both on most evaluation metrics. In Figure 6, we present some examples for qualitatively comparing the segmentation quality between different methods listed in Table 5.

Table 5: Detailed performance contributions of each component in our proposed PLANet.

Methods *		ED			ES		
		D	d_m(mm)	d_H(mm)	D	d_m(mm)	d_H(mm)
$\mathrm{LV_{Endo}}$	BN	0.942±0.028	1.48±0.69	4.89±2.27	0.913±0.042	1.60±0.69	5.12±1.93
	BN+ASPP	0.947±0.023	1.35±0.57	4.53±1.96	0.919±0.041	1.53±0.71	4.81±1.70
	BN+GA	0.948±0.019	1.34±0.48	4.51±1.87	0.920±0.044	1.59±0.68	4.77±1.86
	BN+PLA	0.953±0.020	1.25±0.50	4.39±1.86	0.927±0.035	1.37±0.58	4.42±1.74
	BN+LCL	0.952±0.018	1.26±0.45	4.34±1.65	0.924±0.037	1.47±0.61	4.51±1.77
	PLANet	**0.955±0.016**	**1.21±0.46**	**4.06±1.82**	**0.932±0.029**	**1.30±0.52**	**4.21±1.54**
$\mathrm{LV_{Epi}}$	BN	0.953±0.017	1.78±0.71	5.71±2.54	0.937±0.029	2.14±0.97	5.82±2.18
	BN+ASPP	0.955±0.015	1.71±0.58	5.71±2.39	0.943±0.024	2.08±0.79	5.41±2.06
	BN+GA	0.954±0.014	1.73±0.61	5.32±2.38	0.944±0.026	2.01±0.82	5.21±1.69
	BN+PLA	0.958±0.014	1.66±0.62	**4.99±2.16**	0.946±0.022	1.90±0.78	5.15±1.92
	BN+LCL	0.957±0.013	1.63±0.60	5.05±2.02	0.947±0.024	**1.80±0.80**	5.13±1.82

PLANet	**0.961±0.014**	**1.62±0.62**	5.02±2.34	**0.952±0.022**	1.82±0.71	**4.98±1.71**

* BN: backbone network; ASPP: atrous spatial pyramid pooling; GA: multi-head global attention; PLA: pyramid local attention; LCL: label coherence learning mechanism; PLANet: pyramid local attention network

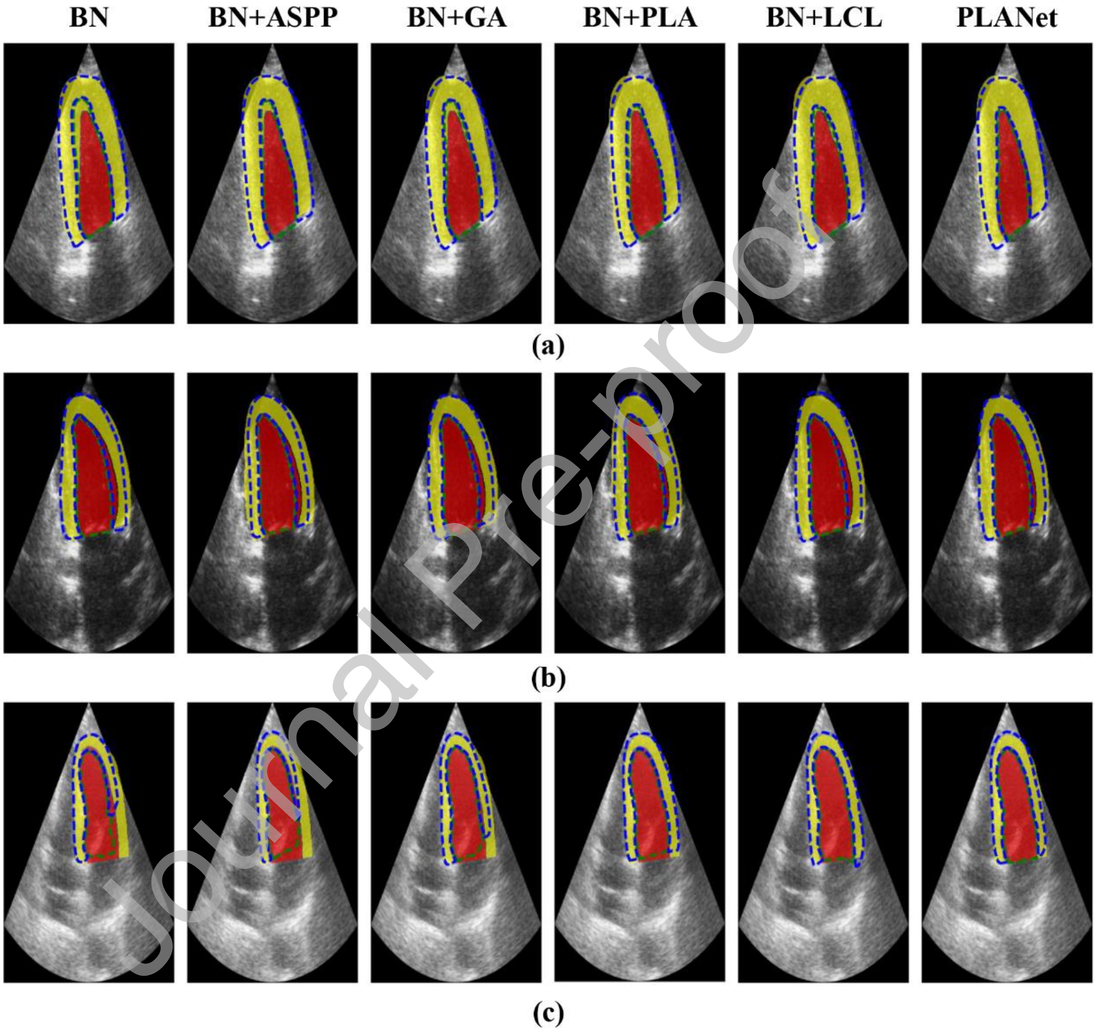

Figure 6: Qualitative comparison between the annotations and segmentation results for different methods. The red and yellow areas represent the annotations of the left ventricle and myocardium, respectively. The green and blue dot lines represent the predicted contours of left ventricle and myocardium, respectively.

3.4 Comparisons with existing methods

We compared the segmentation performance of the proposed PLANet with those of existing methods on the CAMUS and sub-EchoNet-Dynamic datasets. The comparison methods were 1) structured random forests-based method (SRF) (Leclerc et al., 2017); 2) B-spline explicit active surface model (BEASM-fully and BEASM-semi), where the former is fully automatic and the latter is semi-automatic (Pedrosa et al., 2017); 3) U-Net 1 and U-Net 2, where the former is optimized for speed and the latter is optimized for accuracy (Ronneberger et al., 2015); 4) anatomically constrained neural networks (ACNN) (Oktay et al., 2018); 5) stacked hourglasses model (SHG) (Nowell et al., 2016); and 6) U-Net++ (Zhou et al., 2018). The first two are state-of-the-art non-deep learning methods, and the last four are deep learning-based methods. The comparison results of PLANet and the existing methods on the two datasets are presented in Table 6 and Table 7, respectively. The values in boldface represent the best scores for the corresponding metrics. We adopted 10 folds cross-validation for the evaluation on the CAMUS dataset (Leclerc et al., 2019). For sub-EchoNet-Dynamic, we trained PLANet on the training set and evaluated it on the testing set.

The proposed PLANet achieved better segmentation performance for most of the metrics compared with other sophisticated deep learning-based methods and non-deep learning state-of-the-art algorithms. Notably, the segmentation improvement for LV_{Endo} is significant compared to other deep learning-based methods. This is very important because accurate LV_{Endo} segmentation is crucial for the assessment of cardiac functions, which is proven by the outstanding performance on the evaluation of clinical indices in Table 8 and Table 9. Some methods, such as BEASM and ACNN in our experiments, have been integrated with the shape prior to guide the learning of anatomical structures. However, without the prior knowledge, PLANet accurately acquired the knowledge of various cardiac structure shapes by auto-learning from large-scale datasets.

Table 6: Performance comparison of PLANet against existing methods on geometrical metrics on the CAMUS dataset.

Methods		ED			ES		
		D	d_m(mm)	d_H(mm)	D	d_m(mm)	d_H(mm)
LV_{Endo}	SRF	0.895±0.074	2.8±3.6	11.2±10.2	0.848±0.137	3.6±7.8	11.6±13.6
	BEASM-fully	0.879±0.065	3.3±1.8	9.2±4.9	0.826±0.137	3.8±2.1	9.9±5.1
	BEASM-semi	0.920±0.039	2.2±1.2	6.0±2.4	0.861±0.070	3.1±1.6	7.7±3.2
	U-Net 1	0.934±0.042	1.7±1.0	5.5±2.9	0.905±0.063	1.8±1.3	5.7±3.7
	U-Net 2	0.939±0.043	1.6±1.3	5.3±3.6	0.916±0.061	1.6±1.6	5.5±3.8
	ACNN	0.932±0.034	1.7±0.9	5.8±3.1	0.903±0.059	1.9±1.1	6.0±3.9
	SHG	0.934±0.034	1.7±0.9	5.6±2.8	0.906±0.057	1.8±1.1	5.8±3.8
	U-Net++	0.927±0.046	1.8±1.1	6.5±3.9	0.904±0.060	1.8±1.0	6.3±4.2
	PLANet	**0.951±0.018**	**1.3±0.5**	**4.2±1.4**	**0.931±0.032**	**1.4±0.6**	**4.3±1.5**
LV_{Epi}	SRF	0.914±0.057	3.2±2.0	13.0±9.1	0.901±0.078	3.5±4.7	13.0±11.1
	BEASM-fully	0.895±0.051	3.9±2.1	10.6±5.1	0.880±0.054	4.2±2.0	11.2±5.1
	BEASM-semi	0.917±0.038	3.2±1.6	8.2±3.0	0.900±0.042	3.5±1.7	9.2±3.4
	U-Net 1	0.951±0.024	1.9±0.9	5.9±3.4	0.943±0.035	2.0±1.2	6.1±4.1
	U-Net 2	0.954±0.023	1.7±0.9	6.0±3.4	0.945±0.039	1.9±1.2	6.1±4.6
	ACNN	0.950±0.026	1.9±1.1	6.4±4.1	0.942±0.034	2.0±1.2	6.3±4.2
	SHG	0.951±0.023	1.9±1.0	5.7±3.3	0.944±0.034	2.0±1.2	6.0±4.3
	U-Net++	0.945±0.026	2.1±1.0	7.2±4.5	0.939±0.034	2.1±1.1	7.1±5.1
	PLANet	**0.962±0.012**	**1.5±0.5**	**4.6±1.5**	**0.956±0.014**	**1.6±0.6**	**4.6±1.4**

Table 7: Performance comparison of PLANet against existing methods on geometrical metrics on

the sub-EchoNet-Dynamic dataset.

Methods		ED			ES		
		D	d_m(mm)	d_H(mm)	D	d_m(mm)	d_H(mm)
LV_{Endo}	SRF	0.880±0.058	3.4±1.8	11.5±5.8	0.804±0.086	4.8±2.5	14.5±7.2
	BEASM-fully	0.859±0.057	3.4±1.6	10.3±4.3	0.828±0.090	3.6±2.3	10.7±5.6
	BEASM-semi	0.908±0.037	2.2±1.1	8.3±3.3	0.887±0.054	2.3±1.5	8.8±3.5
	U-Net 1	0.931±0.030	1.7±0.8	6.3±3.3	0.906±0.043	1.8±0.9	6.7±3.5
	U-Net 2	0.934±0.026	1.6±0.7	6.1±2.9	0.909±0.39	1.7±0.8	6.6±3.2
	ACNN	0.922±0.031	1.9±0.9	6.7±3.5	0.899±0.053	2.0±1.4	6.9±4.3
	SHG	0.928±0.033	1.9±0.9	6.4±3.5	0.903±0.052	2.1±1.4	6.7±3.8
	U-Net++	0.931±0.030	1.8±0.8	6.7±3.0	0.901±0.042	2.0±1.2	7.2±4.1
	PLANet	**0.942±0.021**	**1.4±0.6**	**5.0±2.2**	**0.918±0.034**	**1.6±0.7**	**5.4±2.6**
LV_{Epi}	SRF	0.893±0.048	3.7±1.7	12.8±5.7	0.879±0.047	4.0±1.9	13.3±6.5
	BEASM-fully	0.906±0.046	3.3±1.7	9.9±4.5	0.893±0.051	3.4±2.0	10.0±5.2
	BEASM-semi	0.911±0.037	2.6±1.4	9.0±3.2	0.906±0.041	2.7±1.7	9.2±5.0
	U-Net 1	0.943±0.027	1.9±0.8	6.5±3.1	0.929±0.031	2.1±1.1	7.1±3.6
	U-Net 2	0.944±0.022	2.0±0.8	7.2±2.8	0.933±0.032	2.0±0.9	7.2±3.2
	ACNN	0.936±0.032	2.3±1.2	7.5±4.0	0.926±0.036	2.3±1.4	7.3±4.4
	SHG	0.937±0.028	2.2±1.0	6.9±3.2	0.930±0.032	2.2±1.3	6.9±3.8
	U-Net++	0.940±0.022	2.1±1.0	7.1±3.7	0.932±0.028	2.2±1.3	7.4±4.6
	PLANet	**0.951±0.017**	**1.7±0.7**	**5.5±2.1**	**0.943±0.019**	**1.8±0.6**	**5.5±2.1**

We estimated LV_{EDV}, LV_{ESV}, and LV_{EF} with the segmentation results of PLANet using the Simpson's rule. The accuracy comparison of different methods is shown in Table 8 and Table 9 for the two datasets. PLANet is able to achieve higher correlations and lower biases than all other methods on most of the metrics. In particular, the correlations between the true LV$_{EF}$ and LV$_{EF}$ estimated based on the segmentation of PLANet were up to 0.883 and 0.869, respectively, which were much higher than the previous best method, U-Net 2. This indicates that with a better segmentation method, 2D echocardiography is able to play a more important role in assessing cardiac functions than is possible at present. In Figures 7 and 8, we present the correlation graphs and Bland Altman analysis for the CAMUS and sub-EchoNet-Dynamic dataset, respectively. The results show a good consistency between the true and predicted clinical indices.

Table 8: Performance comparison of PLANet against existing methods on clinical metrics.

Methods	LV$_{EDV}$		LV$_{ESV}$		LV$_{EF}$	
	corr	*bias*±σ(ml)	*corr*	*bias*±σ(ml)	*corr*	*bias*±σ(%)
SRF	0.755	-0.2±25.7	0.827	9.3±18.0	0.465	11.5±15.4
BEASM-fully	0.704	13.4±30.6	0.713	18.0±25.8	0.731	-9.8±8.3
BEASM-semi	0.886	14.6±19.2	0.880	18.3±16.9	0.790	-9.4±7.2
U-Net 1	0.947	-8.3±12.6	0.955	-4.9±9.9	0.791	**-0.5±7.7**
U-Net 2	0.954	-6.9±11.8	0.964	-3.7±9.0	0.823	-1.0±7.1
ACNN	0.945	-6.7±12.9	0.947	-4.0±10.8	0.799	-0.8±7.5
SHG	0.943	6.4±12.8	0.938	-3.2±11.3	0.770	-1.4±7.8
U-Net++	0.946	-11.4±12.9	0.952	-5.7±10.7	0.789	-1.8±7.7
PLANet	**0.975**	**1.5±8.5**	**0.977**	**0.5±6.3**	**0.882**	0.6±5.8

LV$_{EDV}$: volume of left ventricle in end diastole; LV$_{ESV}$: volume of left ventricle in end systole; *corr*: Pearson correlation coefficient between true and estimated clinical indices; *bias*: mean of bias between true and estimated clinical indices; σ: standard deviation of bias between true and estimated clinical indices

Table 9: Performance comparison of PLANet against existing methods on clinical metrics on sub-EchoNet-Dynamic dataset.

Methods	LV_{EDV}		LV_{ESV}		LV_{EF}	
	corr	*bias*±σ(ml)	*corr*	*bias*±σ(ml)	*corr*	*bias*±σ(%)
SRF	0.945	**-2.3±15.5**	0.943	-14.6±11.8	0.633	15.5±7.9
BEASM-fully	0.960	12.7±15.0	0.956	-4.2±10.2	0.702	13.5±6.2
BEASM-semi	0.978	9.6±10.8	0.976	-0.9±7.4	0.761	7.3±7.3
U-Net 1	0.983	7.2±8.8	0.988	1.2±5.5	0.782	3.1±5.7
U-Net 2	0.981	5.9±9.3	0.988	0.6±5.2	0.827	3.2±5.8
ACNN	0.981	8.8±9.9	0.979	**-0.2±7.0**	0.807	5.8±7.1
SHG	0.972	-3.3±11.2	0.983	-6.6±6.4	0.796	6.0±6.5
U-Net++	0.979	4.9±9.6	0.984	-0.5±5.9	0.819	3.5±5.7
PLANet	**0.986**	2.5±8.6	**0.992**	-0.9±4.6	**0.869**	**2.7±5.9**

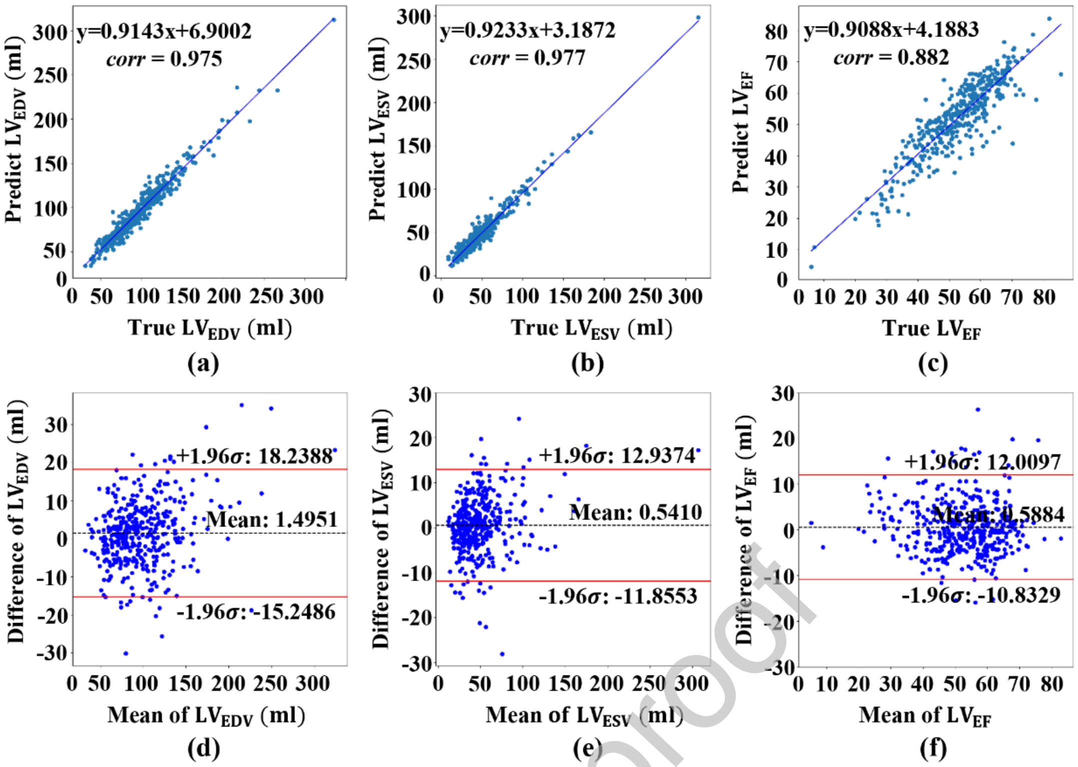

Figure 7: Correlation graphs and Bland Altman graphs for clinical metrics on CAMUS dataset. (a) Correlation between true and predicted LV$_{EDV}$. (b) Correlation between true and predicted LV$_{ESV}$. (c) Correlation between true and predicted LV$_{EF}$. (d) Bland Altman plot between true and predicted LV$_{EDV}$. (e) Bland Altman plot between true and predicted LV$_{ESV}$. (f) Bland Altman plot between true and predicted LV$_{EF}$. In (d)(e)(f), the x-axis represents the mean of true and predicted clinical indices, and the y-axis represents the difference of true and predicted clinical indices.

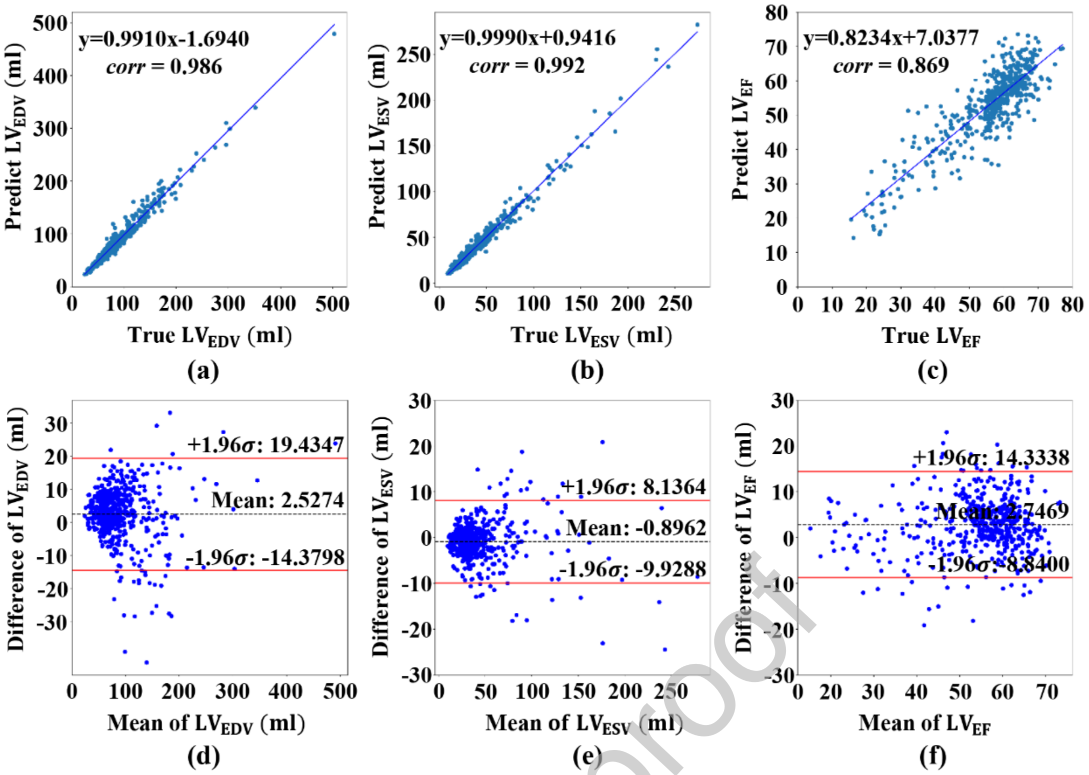

Figure 8: Correlation graphs and Bland Altman graphs for clinical metrics on sub-EchoNet-Dynamic dataset. (a) Correlation between true and predicted LV$_{EDV}$. (b) Correlation between true and predicted LV$_{ESV}$. (c) Correlation between true and predicted LV$_{EF}$. (d) Bland Altman plot between true and predicted LV$_{EDV}$. (e) Bland Altman plot between true and predicted LV$_{ESV}$. (f) Bland Altman plot between true and predicted LV$_{EF}$. In (d)(e)(f), the x-axis represents the mean of true and predicted clinical indices, and the y-axis represents the difference of true and predicted clinical indices.

4. Discussion

In this study, we proposed PLANet for the automatic segmentation of 2D echocardiographic images. Owing to the important role of 2D echocardiography in the diagnosis and treatment of cardiac diseases, PLANet has the potential to assist cardiologists in the overwhelming manual annotation job to save time for more valuable work. We analyzed two unsolved problems in automatic semantic segmentation methods: 1) The lack of an effective feature enhancement approach for capturing supporting features from neighboring contexts, while reducing the influence of image noise. 2) The lack of consideration of label coherence when predicting a category for each pixel to achieve joint-learning across different regions. In this work, we proposed a PLA module and LCL mechanism as our solutions to the two problems, respectively. We evaluated PLANet on CAMUS and sub-EchoNet-Dynamic, two open large-scale datasets in 2D echocardiography, and obtained better segmentation performance compared to existing state-of-the-art methods.

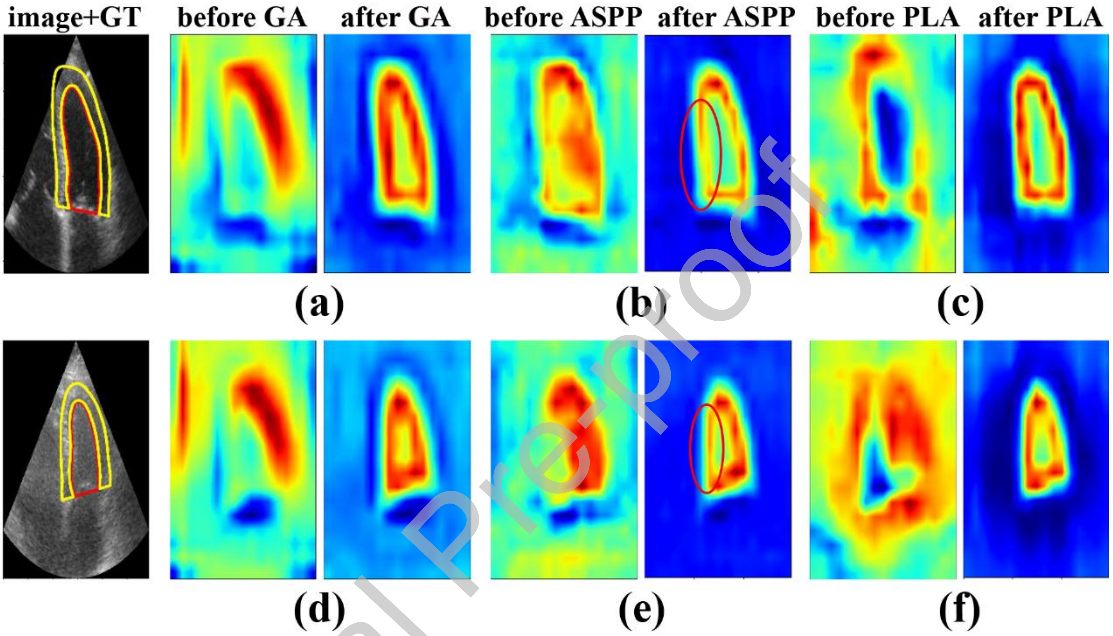

Figure 9: Qualitative comparison of network features for GA, ASPP, and PLA. The first row shows the features of an example with low image contrast between the left ventricle and myocardium. The second row shows the features of an example with lower signal-to-noise ratio. (a)(d) Features before and after GA. (b)(e) Features before and after ASPP. (c)(f) Features before and after PLA. In the first column, the red and yellow lines represent the contours of the left ventricle and myocardium, respectively. GT: ground truth, GA: global attention, ASPP: atrous spatial pyramid pooling.

The PLA module enhances features by aggregating related information from compact and reasonable neighboring contexts. The ablation experiment for PLA proved that a proper size and sparsity of the local context are the requirements for PLA to function effectively. The combination of various sparsities is vital for PLA to collect multi-scale information. A small size limits the effectiveness of context capturing, while a considerably large context size makes the PLA more similar to global attention. Thus, the dispersed attention distribution stacked more image noises, diluting the meaningful features. In Figure 9, we visualized the features of the myocardium before and after the processing of global attention, ASPP, and PLA using the Grad-CAM method (Selvaraju et al., 2019). Global attention, ASPP, and PLA all have a certain ability to enhance semantic features, but differ from each other in some details. Compared with its input, the contextual features of the myocardium after global attention are connected completely, while the features in the border of the myocardium and left ventricle are coarser and have less discrimination (Figures 9a and 9d). This proves the strong ability of the attention mechanism to enhance features by integrating contextual features. However, as the scale of context capturing in global attention is the entire input, supporting information of the local regions is diluted. This results in indecisive and inaccurate semantic features in some key regions such as the contours and apex of the left ventricle. In contrast, the features of ASPP appear heterogeneous in intensity among regions, and the intensities of enhanced features tend to be weaker than the other two (Figures 9b and 9e). The features with low intensity cannot provide a reliable guide to predict for pixels, increasing the probability of mistakes. ASPP avoids coarsening features owing to its local and sparse neighboring feature sampling property. However, unlike PLA, the feature sampling of ASPP is achieved by convolution. As explained in LeCun (1989), weight sharing of convolution can be considered to impose equal constraints among the connection strengths. Therefore, if there is large variance for the conditions of contextual neighboring features among different positions, heterogeneity of semantic features among regions will occur. By contrast, the attention mechanism of PLA makes it possible to adjust connection strengths adaptively for specific contextual conditions. Meanwhile, the localness and sparseness of neighboring context sampling avoid coarsening features by unrelated features and image noises. It can be observed that for the regions with low contrast or high image noise problems, the features across contour are significantly and accurately enhanced (Figures 9c and 9f), which qualitatively proves the effectiveness of the proposed PLA module.

Segmentation prediction by FCN-based semantic segmentation models, without actively considering the coherence across different regions, has been a persistent problem. We proposed a label coherence mechanism to learn the segmentation relations between pixels by offering explicit supervision signals. Our method is easier to implement and train compared to previous approaches. The key point of the LCL mechanism is the supervision signals passed from segmentation label maps as described in Section 2.3. In this work, we proposed a method to parse segmentation label maps flexibly and formulated the training of PLANet as a multi-task of segmentation and label coherence learning. In the ablation experiments, we explored the proper setting of the context scale and sparsity for a fully effective label coherence learning.

The main components of PLANet are the modified backbone network, PLA module and LCL mechanism. We investigated their contributions to the improvement step-by-step in the ablation experiments. The results showed that both PLA and LCL have a positive influence on the performance improvement of semantic segmentation. Moreover, we compared the effectiveness of PLA with ASPP and global attention in the ablation experiment, proving empirically that PLA is better at capturing useful contextual information and enhancing features expressiveness. We presented examples to compare the segmentation results qualitatively in Figure 6. We found that the PLA and LCL mechanism are especially helpful to the segmentation on apex of the left ventricle (Figure 6a). This is important to accurately locate the vertical axis when computing LV_{EDV} and LV_{ESV} with the Simpson's rule. PLANet was also better than other methods when segmenting on images with low contrast (Figure 6b) and high noises (Figure 6c).

Compared to the existing semantic segmentation methods, the proposed PLANet performed better on most of the evaluation metrics. From the geometrical perspective, the segmentation improvement on LV_{Endo} were larger than that on LV_{Epi}, which is important because the low contrast between LV_{Endo} and LV_{Epi} has hindered the robustness of segmentation methods for long (Leclerc et al., 2019). Segmentation improvement on LV_{Epi} may be relatively smaller because the Dice of LV_{Epi} tends toward saturation. From the clinical perspective, the *corr*s of LV_{EF} were improved up to 0.882 and 0.869 on the two datasets, which are great improvements compared to all the existing methods. This shows that the reliability of volumetric measurements on 2D echocardiographic images can be higher if a more effective segmentation method is available.

The current version of PLANet has limitations. The memory-consumption of the network training is still large, which can cause memory explosion in the training phase. We proposed a simple scheme to balance the proper batch size and memory-consumption in Section 2.4, but this prolongs training to some extent. Further work is required in the future to explore a more proper method of PLANet implementation.

5. Conclusion

In this work, we proposed a novel PLANet method for the semantic segmentation of 2D echocardiographic images. With the application of the PLA module and LCL mechanism, our method achieves accurate and efficient automatic segmentation for multiple cardiac structures, indicating the potential of PLANet as an assistive tool in clinical practice.

Credit Author Statement

Fei Liu: Data curation, Resources, Conceptualization, Methodology, Software, Formal analysis, Writing-Original draft preparation, Writing-Reviewing and Editing. **Kun Wang:** Conceptualization, Writing-Reviewing and Editing, Supervision. **Dan Liu:** Data curation, Resources. **Xin Yang:**

Investigation, Validation. **Jie Tian:** Conceptualization, Funding acquisition, Project administration, Supervision.

Declaration of interests

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

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests

Acknowledgements

This study is supported by Ministry of Science and Technology of China under Grant No. 2017YFA0205200, National Natural Science Foundation of China under Grant Nos. 81930053, 61671449, 81227901, and 81527805, Chinese Academy of Sciences under Grant Nos. KFJ-STS-ZDTP-059, YJKYYQ20180048, QYZDJ-SSW-JSC005, and XDB32030200. The authors would like to acknowledge Dr. Dan Liu and her team for their outstanding contributions in building the sub-EchoNet-Dynamic dataset.

References

Al-Barazanchi, H.A., Qassim, H., Verma, A., 2016. Novel CNN architecture with residual learning and deep supervision for large-scale scene image categorization. IEEE Annu. Ubiquitous Comput., Electron. Mob. Commun. Conf. pp. 1-7. https://doi.org/10.1109/UEMCON.2016.7777858

Araújo, T., Aresta, G., Mendonça, L., Penas, S., Maia, C., Carneiro, Â., Mendonça, A.M., Campilho, A., 2020. DR|GRADUATE: Uncertainty-aware deep learning-based diabetic retinopathy grading in eye fundus images. Med. Image Anal. 63, 101715. https://doi.org/10.1016/j.media.2020.101715

Bian, C., Yuan, C., Wang, J., Li, M., Yang, X., Yu, S., Ma, K., Yuan, J., Zheng, Y., 2020. Uncertainty-aware domain alignment for anatomical structure segmentation. Med. Image Anal. 64, 101732. https://doi.org/10.1016/j.media.2020.101732

Carneiro, G., Nascimento, J., Freitas, A., 2012. The segmentation of the left ventricle of the heart from ultrasound data using deep learning architectures and derivative-based search methods. IEEE Trans. Image Process. 21, pp. 968-982. https://doi.org/10.1109/TIP.2011.2169273

Carneiro, G., Nascimento, J.C., 2012. The use of on-line co-training to reduce the training set size in pattern recognition methods: Application to left ventricle segmentation in ultrasound. Proc. IEEE Comput. Vis. Pattern Recognit. pp. 948-955. https://doi.org/10.1109/CVPR.2012.6247770

Carneiro, G., Nascimento, J.C., 2013. Combining multiple dynamic models and deep learning architectures for tracking the left ventricle endocardium in ultrasound data. IEEE Trans. Pattern Anal. Mach. Intell. 35, pp. 2592-2607. https://doi.org/10.1109/TPAMI.2013.96

Chen, C., Qin, C., Qiu, H., Tarroni, G., Duan, J., Bai, W., Rueckert, D., 2020. Deep learning for cardiac image segmentation: A review. Front. Cardiovasc. Med. 7, 25. https://doi.org/10.3389/fcvm.2020.00025

Chen, H., Zheng, Y., Park, J.-H., Heng, P.-A., Zhou, S.K., 2016. Iterative multi-domain regularized deep learning for anatomical structure detection and segmentation from ultrasound images. Proc. Med. Image Comput. Comput. Assist. Interv. pp. 487-495. https://doi.org/10.1007/978-3-319-46723-8_56

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2018. Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans. Pattern Anal. Mach. Intell. 40, pp. 834-848. https://doi.org/10.1109/TPAMI.2017.2699184

Chen, L.-C., Papandreou, G., Schroff, F., Adam, H., 2017. Rethinking atrous convolution for semantic image segmentation. Proc. IEEE Comput. Vis. Pattern Recognit. arXiv:1706.05587

Crewther, D.P., Lawson, M.L., Crewther, S.G., 2007. Global and local attention in the attentional blink. J.

Vis. 7, pp. 1-12. https://doi.org/10.1167/7.14.9

Davatzikos, C., Tao, X., Shen, D., 2003. Hierarchical active shape models, using the wavelet transform. IEEE Trans. Med. Imaging. 22, pp. 414-423. https://doi.org/10.1109/TMI.2003.809688

Dong, S., Luo, G., Tam, C., Wang, W., Wang, K., Cao, S., Chen, B., Henggui, Z., Li, S., 2020. Deep atlas network for efficient 3D left ventricle segmentation on echocardiography. Med. Image Anal. 61, 101638. https://doi.org/10.1016/j.media.2020.101638

Gregor, K., Danihelka, I., Graves, A., Rezende, D.J., Wierstra, D., 2015. DRAW: A recurrent neural network for image generation. Proc. Int. Conf. Mach. Learn. pp. 1462-1471. arXiv:1502.04623

Guo, X., Yuan, Y., 2020. Semi-supervised WCE image classification with adaptive aggregated attention. Med. Image Anal. 64, 101733. https://doi.org/10.1016/j.media.2020.101733

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. pp. 770-778. https://doi.org/10.1109/CVPR.2016.90

Heinze, H.-J., Münte, T.F., 1993. Electrophysiological correlates of hierarchical stimulus processing: Dissociation between onset and later stages of global and local target processing. Neuropsychologia. 31, pp. 841-852. https://doi.org/10.1016/0028-3932(93)90132-J

Hu, J., Shen, L., Sun, G., 2019. Squeeze-and-excitation networks. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. pp. 7132-7141. https://doi.org/10.1109/CVPR.2018.00745

Huang, X., Dione, D.P., Compas, C.B., Papademetris, X., Lin, B.A., Bregasi, A., Sinusas, A.J., Staib, L.H., Duncan, J.S., 2014. Contour tracking in echocardiographic sequences via sparse representation and dictionary learning. Med. Image Anal. 18, pp. 253-271. https://doi.org/10.1016/j.media.2013.10.012

Jafari, M.H., Girgis, H., Liao, Z., Behnami, D., Abdi, A., Vaseli, H., Luong, C., Rohling, R., Gin, K., Tsang, T., Abolmaesumi, P., 2018. A unified framework integrating recurrent fully-convolutional networks and optical flow for segmentation of the left ventricle in echocardiography data. Deep Learn. Med.l Imaging Anal. Multimodal Learn. Clin. Decis. pp. 29-37. https://doi.org/10.1007/978-3-030-00889-5_4

Khosravan, N., Celik, H., Turkbey, B., Jones, E.C., Wood, B., Bagci, U., 2019. A collaborative computer aided diagnosis (C-CAD) system with eye-tracking, sparse attentional model, and deep learning. Med. Image Anal. 51, pp. 101-115. https://doi.org/10.1016/j.media.2018.10.010

Leclerc, S., Grenier, T., Espinosa, F., Bernard, O., 2017. A fully automatic and multi-structural segmentation of the left ventricle and the myocardium on highly heterogeneous 2D echocardiographic data. Proc. IEEE Ultrason. Symp. pp. 1-4. https://doi.org/10.1109/ULTSYM.2017.8092797

Leclerc, S., Smistad, E., Pedrosa, J., Østvik, A., Cervenansky, F., Espinosa, F., Espeland, T., Rye Berg, E.A., Jodoin, P.-M., Grenier, T., Lartizien, C., D'hooge, J., Lovstakken, L., Bernard, O., 2019. Deep learning for segmentation using an open large-scale dataset in 2D echocardiography. IEEE Trans. Med. Imaging. 38, pp. 2198-2210. https://doi.org/10.1109/TMI.2019.2900516

LeCun, Y., 1989. Generalization and network design strategies. Connectionism in Perspective.

Lei, B., Huang, S., Li, H., Li, R., Bian, C., Chou, Y.-H., Qin, J., Zhou, P., Gong, X., Cheng, J.-Z., 2020a. Self-co-attention neural network for anatomy segmentation in whole breast ultrasound. Med. Image Anal. 64, 101753. https://doi.org/10.1016/j.media.2020.101753

Lei, Y., Tian, Y., Shan, H., Zhang, J., Wang, G., Kalra, M.K., 2020b. Shape and margin-aware lung nodule classification in low-dose CT images via soft activation mapping. Med. Image Anal. 60, 101628. https://doi.org/10.1016/j.media.2019.101628

Li, L., Tang. S., Zhang, Y., Deng, L., Tian, Q., 2018. GLA: Global-local attention for image description. IEEE Trans. Multimed. 20, pp. 726-737. https://doi.org/10.1109/TMM.2017.2751140

Liu, Z., Li, X., Luo, P., Loy, C.C., Tang, X., 2018. Deep learning markov random field for semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 40, pp. 1814-1828. https://doi.org/10.1109/TPAMI.2017.2737535

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. IEEE. https://doi.org/10.1109/CVPR.2015.7298965

Luong, M.-T., Pham, H., Manning, C.D., 2015. Effective approaches to attention-based neural machine translation. Proc. 2015 Conf. Empir. Methods Nat. Ling. pp. 1412-1421. https://doi.org/10.18653/v1/D15-1166

Maicas, G., Bradley, A.P., Nascimento, J.C., Reid, I., Carneiro, G., 2019. Pre and post-hoc diagnosis and interpretation of malignancy from breast DCE-MRI. Med. Image Anal. 58, 101562. https://doi.org/10.1016/j.media.2019.101562

Mirsamadi, S., Barsoum, E., Zhang, C., 2017. Automatic speech emotion recognition using recurrent neural networks with loal attention. Proc. IEEE Int. Conf. Acoust. Speech Signal Process. pp. 2227-2231. https://doi.org/10.1109/ICASSP.2017.7952552

Nascimento, J.c., Carneiro, G., 2019. One shot segmentation: unifying rigid detection and non-rigid segmentation using elastic regularization. IEEE Trans. Pattern Anal. Mach. Intell. https://doi.org/10.1109/TPAMI.2019.2922959

Newell, A., Yang, K., Deng, J., 2016. Stacked hourglass networks for human pose estimaiton. Proc. Eur. Conf. Comput. Vis. pp. 483-499. https://doi.org/10.1007/978-3-319-46484-8_29

Oktay, O., Ferrante, E., Kamnitsas, K., Heinrich, M., Bai, W., Caballero, J., Cook, S.A., De Marvao, A., Dawes, T., O'Regan, D.P., Kainz, B., Glocker, B., Ruekert, D., 2018. Anatomically constrained neural networks (ACNNs): Applicaiton to cardiac image enhancement and segmentation. IEEE Transact. Med. Imaging. 37, pp. 384-395. https://doi.org/10.1109/TMI.2017.2743464

Oktay, O., Shi, W., Keraudren, K., Caballero, J., Rueckert, D., Hajnal, J., 2014. Learning shape representations for multi-atlas endocardium segmentation in 3D echo images. Proc. Med. Image Comput. Comput. Assist. Interv. pp. 57-64. https://doi.org/10.13140/2.1.3767.5522

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., Zou, J.Y., 2020. Video-based AI for beat-to-beat assessment of cardiac function. Nature 580, pp. 252-256. https://doi.org/10.1038/s41586-020-2145-8

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., Devito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A., 2017. Automatic differentiation in PyTorch. Proc. Adv. Neural Inf. Process. Syst. 1-4.

Pedrosa, J., Queirós, S., Bernard, O., Engvall, J., Edvardsen, T., Nagel, E., D'hooge, J., 2017. Fast and fully automatic left ventrilar segmentation and tracking in echocardiography using shape-based b-spline explicit active surfaces. IEEE Trans. Med. Imaging. 36, pp. 2287-2296. https://doi/org/10.1109/TMI.2017.2734959

Peng, C., Zhang, X., Yu, G., Luo, G., Sun, J., 2017. Large kernel matters -- Improve semantic segmentation by global convolutional network. Proc. IEEE Comput. Vis. Pattern Recognit. pp. 4353-4361. https://.doi.org/10.1109/CVPR.2017.189

Pesce, E., Withey, S.J., Ypsilantis, P.-P., Bakewell, R., Goh, V., Montana, G., 2019. Learning to detect chest radiographs containing pulmonary lesions using visual attention networks. Med. Image Anal. 53, 26-38. https://doi.org/10.1016/j.media.2018.12.007

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutonal networks for biomedical image segmentation. Proc. Med. Image Comput. Comput. Assist. Interv. pp. 234-241. https://doi.org/10.1007/978-3-319-24574-4_28

Selvaraju, R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2019. Grad-CAM: Visual explanations from deep networks via gradient-based localization. Proc. IEEE Int. J. Comput. Vis. 128, pp.336-359. https://doi.org/10.1007/s11263-019-01228-7

Shelhamer, E., Long, J., Darrell, T., 2017. Fully convolutional networks for semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39, pp. 640-651. https://doi.org/10.1109/TPAMI.2016.2572683

Shi, S.-Y., Lee, S., Yun, I.-D., Lee, K.-M., 2019. Deep vessel segmentation by learning graphical connectivity. Med. Image Anal. 58, 101556. https://doi.org/10.1016/j.media.2019.101556

Smistad, E., Lindseth, F., 2014. Real-time tracking of the left ventricle in 3D ultrasound using kalman filter and mean value coordinates. Proc. Med. Image Comput. Comput. Assist. Interv. pp. 65-72. https://doi.org/10.13140/2.1.1330.6888

Stolte, S., Fang, R., 2020. A survey on medical image analysis in diabetic retinopathy. Med. Image Anal. 64, 101742. https://doi.org/10.1016/j.media.2020.101742

Taghanaki, S.A., Abhishek, K., Cohen, J.P., Cohen-Adad, J., Hamarneh, G., 2020. Deep semantic segmentation of natural and medical images: A review. Artif. Intell. Rev. pp. 1-42 https://doi.org/10.0462-020-09854-1

van Stralen, M., Haak, A., Leung, K.Y.E., van Burken, G., Bosch, J.G., 2014. Segmentation of multi-center 3D left ventricular echocardiograms by active appearance models. Proc. Med. Image Comput. Comput. Assist. Interv. pp. 57-64.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. Proc. Adv. Neural Inf. Process. Syst. pp. 5998-6008. arXiv:1706.03762

Veni, G., Moradi, M., Bulu, H., Narayan, G., Syeda-Mahmood, T., 2018. Echocardiography segmentation based on a shape-guided deformable model driven by a fully convolutional network prior. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recogn. pp. 898-902. https://doi.org/10.1109/ISBI.2018.8363716

Wang, C., Smedby, Ö., 2014. Model-based left ventricle segmentation in 3D ultrasound using phase image. Proc. Med. Image Comput. Comput. Assist. Interv. pp. 81-88.

Wang, L., Zhang, L., Zhu, M., Qi, X., Yi, Z., 2020. Automatic diagnosis for thyroid nodules in ultrasound images by deep neural networks. Med. Image Anal. 61, 101665. https://doi.org/10.1016/j.media.2020.101665

Wang, S., Yaxi, Z., Yu, L., Chen, H., Lin, H., Wan, X., Fan, X., Heng, P.-A., 2019a. RMDL: Recalibrated multi-instance deep learning for whole slide gastric image classification. Med. Image Anal. 58, 101549. https://doi.org/10.1016/j.media.2019.101549

Wang, Y., Deng, Z., Hu, X., Zhu, L., Yang, X., Xu, X., Heng, P.-A., Ni, D., 2018. Deep attentional features for prostate segmentation in ultrasound. Proc. Med. Image Comput. Comput. Assist. Interv. pp. 523-530. https://doi.org/10.1007/978-3-030-00937-3_60

Wang, Y., Zhou, Y., Shen, W., Park, S., Fishman, E.K., Yuille, A.L., 2019b. Abdominal multi-organ segmentation with organ-attention networks and statistical fusion. Med. Image Anal. 55, 88-102. https://doi.org/10.1016/j.media.2019.04.005

Wei, W., Poirion, E., Bodini, B., Durrleman, S., Ayache, N., Stankoff, B., Colliot, O., 2019. Predicting PET-derived demyelination from multimodal MRI using sketcher-refiner adversarial training for multiple sclerosis. Med. Image Anal. 58, 101546. https://doi.org/10.1016/j.media.2019.101546

Xia, B.N., Gong, Y., Zhang, Y., Poellabauer, C., 2019. Second-order non-local attention networks for person re-identification. Proc. IEEE Int. Conf. Comput. Vis. 3760-3769. https://doi.org/10.1109/ICCV.2019.00386

Xu, C., Xu, L., Ohorodnyk, P., Roth, M., Chen, B., Li, S., 2020. Contrast agent-free synthesis and segmentation of ischemic heart disease images using progressive sequential causal GANs. Med. Image Anal. 62, 101668. https://doi.org/10.1016/j.media.2020.101668

Yu, C., Wang, J., Peng, C., Goao, C., Yu, G., Sang, N., 2018. BiSeNet: Bilateral segmentation network for real-time semantic segmentation. Proc. Eur. Conf. Comput. Vis. pp. 334-349. arXiv:1808.00897

Yu, L., Guo, Y., Wang, Y., Yu, J., Chen, P., 2017. Segmentation of fetal left ventricle in echocardiographic sequences based on dynamic convolutional neural networks. IEEE Trans. Biomed. Eng. 64, pp. 1886-1895. https://doi.org/10.1109/TBME.2016.2628401

Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., 2015. Conditional random fields as recurrent neural networks, Proc. IEEE Int. Conf. Comput. Vis. pp. 1529-1537. https://doi.org/10.1109/ICCV.2015.179

Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J., 2018. UNet++: A nested u-net architecture for medical image segmentation. Lect. Notes Comput. Sci. pp. 3-11. https://doi.org/10.1007/978-3-030-00889-5_1