Applied Soft Computing Journal 88 (2020) 106049

Contents lists available at ScienceDirect

Applied Soft Computing Journal

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

Unified model for interpreting multi-view echocardiographic sequences without temporal information

Ming Li [a,b], Shizhou Dong [a,b], Zhifan Gao [c], Cheng Feng [d], Huahua Xiong [e], Wei Zheng [a], Dhanjoo Ghista [f], Heye Zhang [g,*], Victor Hugo C. de Albuquerque [h]

[a] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
[b] Shenzhen College of Advanced Technology, University of Chinese Academy of Sciences, Shenzhen, China
[c] Department of Medical Imaging, Western University, London ON, Canada
[d] Department of Ultrasound, The Third People's Hospital of Shenzhen, Shenzhen, China
[e] Department of Ultrasound, The First Affiliated Hospital of Shenzhen University, Shenzhen, China
[f] University 2020 Foundation, United States
[g] School of Biomedical Engineering, Sun Yat-sen University, Shenzhen, China
[h] University of Fortaleza, Fortaleza, Brazil

ARTICLE INFO

Article history:
Received 9 May 2019
Received in revised form 20 December 2019
Accepted 21 December 2019
Available online 7 January 2020

Keywords:
Unified model
Dense pyramid
Deep supervision
Multi-view
Echocardiographic sequence
Multi-vendor
Multi-center
Without temporal information

ABSTRACT

The robust and fully automatic interpretation of multi-view echocardiographic sequences across multi-vendor and multi-center is a challenging task due to abounding artifacts, low signal-to-noise ratio, large shape variations among different views, and large gaps across different centers and vendors. In this paper, a dense pyramid and deep supervision network (DPSN) is proposed to tackle this challenging task. DPSN incorporates the advantages of the densely connected network, feature pyramid network, and deeply supervised network, which help to extract and fuse multi-level and multi-scale holistic semantic information. This capability endows DPSN with prominent generalization and robustness, enabling it to yield a precise interpretation. To reduce the computational complexity and avoid the frequent information loss in temporal modeling, DPSN processes all frames independently (i.e., without utilizing temporal information) but can still obtain stable and coherent performance in the sequence. Adequate experiments on the heterogeneous (multi-view, multi-center, and multi-vendor) dataset (10858 labeled images) corroborate that DPSN achieves not only superior segmentation results but also prominent computational efficiency and stable performance. Estimation of the ejection fraction also shows good clinical correlation, revealing the clinical potential of DPSN.

1. Introduction

Multi-view echocardiographic sequences interpretation provides important insights into the prognosis and diagnosis of cardiac diseases. The knowledge pattern of cardiac structures and textures associated with deforming tissues can be observed in echocardiographic sequences whereas in single frames the information is always missing and incomplete [1,2]. Echocardiographic sequences also permit continuous observation of the heart, which is useful for assessing the wall [3] and for identifying the end-diastolic (ED) and end-systolic (ES) phases [4]. Cardiologists usually check multi-view echocardiographic sequences to obtain a more comprehensive functional assessment of the heart in clinical decision making [5,6]. The apical-2-chamber view (A2C), apical-3-chamber view (A3C), and apical-4-chamber view (A4C) are the most commonly used views for functional assessment of the left ventricle (LV) [7]. A2C is useful for assessing the anterior and inferior walls of the LV [8]; A3C helps in assessing the contractility of the anterolateral and posterior walls [9]; and A4C is the best view for visualizing the apex of the LV, which is utilized to assess the overall function of the LV and the contractility of the inter-ventricular septum, apex and lateral walls [3,10]. Most clinical indices of the LV (e.g., area, volume, and ejection fraction) are measured using these three standard apical views [7]. Segmentation of the LV is generally a prerequisite for such quantitative analysis of clinical indices. For instance, the biplane method of disks summation (modified Simpson's rule), which is the most commonly used method for the LV volume calculation, requires accurate segmentation of the LV in A2C and A4C [7].

In clinical practice, quantitative analysis of the LV still involves careful review and laborious manual interpretation by experts, which are tedious and time-consuming. Thus, automatic methods are desired to facilitate this process. However, multi-view

* Corresponding author.
E-mail address: zhangheye@mail.sysu.edu.cn (H. Zhang).

https://doi.org/10.1016/j.asoc.2019.106049
1568-4946/

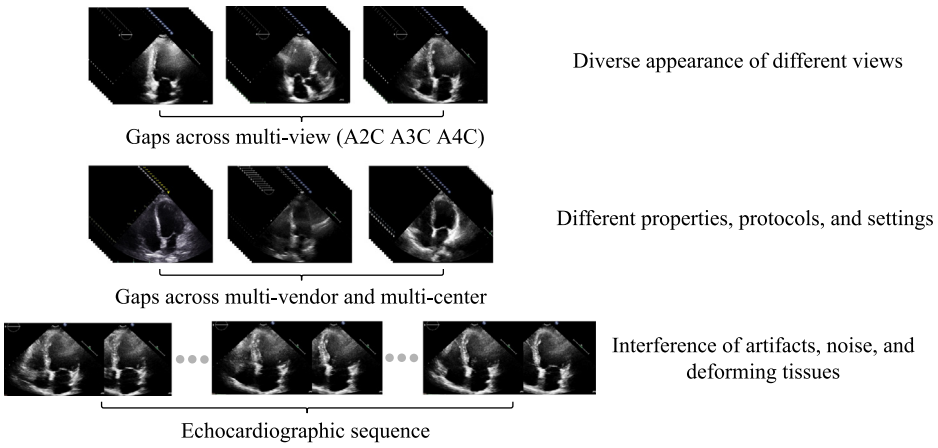

Fig. 1. Top row: multi-view samples; Middle row: A4C samples across centers and vendors; Bottom row: echocardiographic sequence.

echocardiographic sequences segmentation remains a challenging task due to the following challenges and problems (also as illustrated in Fig. 1):

- Image quality: the fuzzy border definition caused by the presence of noise and artifacts, edge dropout, low signal-to-noise ratio, high variability, and poor contrast, resulting in local missing and incomplete information of the anatomical structure [1].
- Multi-view: different views have different anatomical structures, leading to the quite varied appearance of the LV. They provide complementary information in clinical diagnosis but also make it difficult to establish uniform semantic features for a generalized and robust model.
- Multi-center and multi-vendor: under different specific properties, protocols, and settings of different centers and vendors, there will be considerable variations in the gray value distribution and spatial texture of images, resulting in large gaps across centers and vendors. These gaps hinder the extraction and fusion of holistic semantic information, thus impeding the development of a generalized and robust model.
- Echocardiographic sequences: compared with the ED and ES frames, other frames in the cardiac cycle suffer from additional interference of the unclosed mitral valve, endocardial trabeculation, and papillary muscles, causing extensive blurring in the borders of the LV [11].

Application scenarios for existing methods are always limited, making them suitable only in specific situations. These methods mostly focus on a single view, on specific frames (ED and ES), or on a single vendor and center [11–13]. For sequence segmentation, existing methods try to leverage temporal information by using a deformable model combined with optical flow [14–16] or recurrent neural network (RNN) [17,18]. The major downsides of these temporal methods are that they are computationally cumbersome and suffer from information loss. As yet, there is still no study that achieves a unified model with generalization and robustness to accommodate heterogeneous data efficiently.

To settle the aforementioned challenges, a dense pyramid and deep supervision network is presented in this paper. The workflow of DPSN is depicted in Fig. 2. The extraction network extracts multi-level and multi-scale holistic semantic features, endowing DPSN with the superior feature extraction ability and the LV region detection capacity in multi-level and multi-scale space. Next, the fusion network fuses multi-level and multi-scale holistic semantic features, helping to constrain the LV boundaries and provide better semantic representation. The forward flow refines the segmentation results, while the backward flow regularizes the feature extraction by deep supervision. They exert mutual promotion, helping DPSN to capture the global geometric characteristic of the LV and establish uniform semantic features. To reduce the computational complexity and avoid frequent information loss in temporal modeling, DPSN is designed to process each frame independently without utilizing temporal information. Ultimately, DPSN can achieve a unified model with generalization and robustness to accommodate heterogeneous data, not only generating accurate segmentation results but also obtaining prominent computational efficiency and stable performance in the sequence.

2. Literature review

Active contours, level sets, active shape models, Kalman filter and many other traditional methods have been proposed for LV segmentation; a survey of these methods was summarized by Noble and Boukerroui in [1]. Recently, with the rapid growth of computational power, deep learning methods have become prevalent in natural image classification [19,20] and segmentation [21,22] tasks, as well as for medical image analysis [23,24].

2.1. Existing methods for specific frames and views

Convolutional neural networks (CNN) based methods have outperformed hand-engineered methods in medical image segmentation tasks [25,26]. High-level features learned from CNN achieve better performance than low-level hand-crafted features because low-level features lacking contextual semantic information that high-level features have. Deep belief networks (DBN) combined with a derivative-based search strategy were used to identify the LV region of interest and adapt a spline to edges [27]. However, DBN has to start a new complex search for each view and each sequence, which is computationally cumbersome. Oktay *et al.* proposed ACNN [28] to incorporate anatomical prior knowledge into CNN for LV segmentation, but they focused only on ED and ES frames segmentation. Fully convolutional networks (FCN) have performed excellently on image segmentation tasks [21]. Chen *et al.* proposed a multi-domain regularized fully convolutional network (iMD-FCN) built upon FCN for multi-view LV segmentation [29]. However, iMD-FCN needs to detect and crop the LV region iteratively, and it cannot handle multiple frames segmentation in the cardiac cycle. By adding skip connections between the contracting and expanding paths, U-Net showed prominent segmentation accuracy in medical image segmentation tasks [23,30,31]. Zhang *et al.* utilized U-Net for multi-view

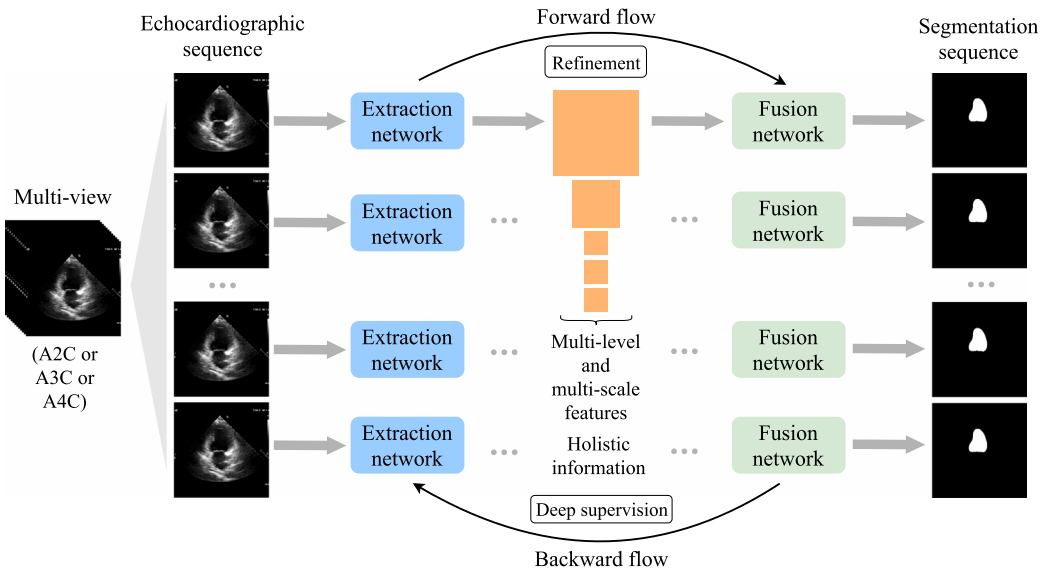

Fig. 2. Overview of the workflow. The extraction network and fusion network constitute an end-to-end framework, and they exert mutual promotion and help to achieve a unified model with generalization and robustness to accommodate heterogeneous data.

LV segmentation, but they trained multiple models for different views indirectly, rather than achieving a generalized and robust model directly [32].

DPSN differs from typical methods by focusing on the direct segmentation of multi-view echocardiographic sequences using one unified model, which is a more complicated end-to-end application. Since a single model has been shown to be effective in diagnosis across medical modalities [24], it is possible to achieve a generalized and robust unified model to accommodate heterogeneous data. To achieve such a generalized and robust model that covers sufficient variability, a dense pyramid and deep supervision network is developed to capture the global geometric characteristic of the LV. The workflow of DPSN is illustrated in Fig. 2. Unlike the indirect operations of typical methods, which need to segment the LV separately after network retraining when applied to other views [32,33], or detect and crop the LV region iteratively [29], DPSN can achieve a unified model with generalization and robustness for direct multi-view echocardiographic sequences segmentation.

2.2. Existing methods for sequence

Regarding sequence segmentation, typical methods focus on temporal modeling, assuming that the target anatomical structure does not vary sharply between consecutive frames [2,11,15,34]. They search for targets whose shape and appearance vary gradually in consecutive frames; however, these constraints may fail to recover from relatively common situations, such as occlusions of abrupt changes or large noise [35,36]. Furthermore, the exploitation of temporal modeling is not a trivial task, typical methods have to compute temporal matches in the form of optical flow [14,16] or RNN [17,18], both of which are computationally cumbersome and inefficient. Several RNN based methods [30, 37,38] have been implemented for MRI sequence segmentation. Although they work well in MRI, they may not be applicable to echocardiography. Because the variations in anatomical structures, noise, and disturbances of deforming tissues in echocardiography are much more severe than those in MRI, and human eye compensation is usually required to resolve ambiguities and determine the exact LV boundary location [39]. One major problem with existing RNN based methods is that they are too sensitive to noise disturbances and insufficiently robust; consequently, even slight disturbances can cause large changes in subsequent results. The loop in RNN amplifies the noise and disturbs the representation learning [40,41].

In contrast to existing methods that focus on temporal modeling with explicitly imposed and expensive constraints, DPSN casts the cardiac cycle segmentation as an independent frame segmentation task. Given the different extents of noise, especially when noise temporarily covers a large section of the LV anatomical structure due to improper operation by the operator (i.e., parts of the LV anatomical structure may be missing in some frames), it is difficult to establish temporal correlation for echocardiographic sequences. Occlusions of abrupt motion or severe noise (e.g., motion of the mitral valve) cause the shape and appearance to vary acutely in consecutive frames, resulting in the frequent information loss in temporal modeling; hence, temporal constraints fail to recover from these situations. Moreover, due to the strong constraints of temporal modeling, even a small error caused by noise will result in a series of deviations, propagating errors to the subsequent frames in the cardiac cycle and causing considerable temporal error accumulation.

Therefore, in the cardiac cycle segmentation task, time continuity is not a problem that must be considered. Rather, more attention should be paid to improve the accuracy and robustness of the LV sequence segmentation. Thus, DPSN is designed to process each frame independently in the cardiac cycle. This procedure has some natural advantages: more robust to noise disturbances and occlusions; not limited to certain ranges of motion and constraints; frames need not be processed sequentially, and errors will not be propagated temporally. In practice, these allow DPSN to handle noisy frames in which the LV anatomical structures can be partially missing. Without temporal modeling, there will be no temporal error accumulation and frequent information loss; rather, much better accuracy and a faster process will be achieved.

2.3. Contributions

The primary contributions of the proposed method are summarized as follows:

- For the first time, a unified model with generalization and robustness to accommodate heterogeneous data efficiently for echocardiographic sequence segmentation is achieved.

- A dense pyramid and deep supervision network is presented to incorporate and extend the advantages of the densely connected network, feature pyramid network, and deeply supervised network into an end-to-end framework. This framework is powerful in the extraction and fusion of multi-level and multi-scale holistic semantic information.
- DPSN processes all frames independently without utilizing temporal information, but it can still obtain stable and coherent performance in the sequence. DPSN not only reduces the computational complexity, but also eliminates temporal error accumulation and avoids the frequent information loss in temporal modeling.

The rest of this paper is organized as follows: Section 3 describes the details of the proposed method; Section 4 provides adequate experimental results to demonstrate the robustness and accuracy of DPSN; Section 5 draws some conclusions.

3. Methodology

DPSN incorporates the insights of densely connected network (DenseNet) [42], feature pyramid network (FPN) [43], and deeply supervised network [44] to those of plain U-Net to extract and fuse multi-level and multi-scale holistic semantic information. It is built as an end-to-end framework and comprised of three key components: the deep feature pyramid module, the pyramid pooling module, and the deep supervision module (as depicted in Fig. 3). The deep feature pyramid module consists of pyramidal dilated dense convolution blocks (ConvBlocks). The pyramid pooling module contains parallel poolings and convolutions. While the deep supervision module involves a series of composite losses.

3.1. Multi-level and multi-scale features extraction

The deep feature pyramid module is designed as a deep pyramid hierarchical architecture. It includes 5 levels of ConvBlocks to extract multi-level and multi-scale holistic semantic features. Multi-level information captures the global geometric characteristics of the LV, while multi-scale information strengthens thin and small regions, helping to refine the boundaries of the LV. They contribute to lessening the gap among views, centers, and vendors, which increases the robustness of DPSN to image conditions and anatomical structure variations.

Multi-level and multi-scale features contain both low-level detailed information and high-level semantic information. To fuse low-level features and high-level features and bring awareness of multi-scale features, inspired by FPN, the feature pyramid representation architecture is developed with predictions be made independently from different level feature maps and connected laterally to address multi-scale problems. Five ConvBlocks are integrated into different abstraction levels as the main module, which is inspired by DenseNet. One ConvBlock contains L densely connected dilated convolution layers, as shown in Fig. 4, which can expand the receptive field and meanwhile preserve the resolution of feature maps. While the transition layer changes the channels and resolution of feature maps by convolution and pooling. The feedforward information propagation from the preceding l layers to the $(l+1)$th layer can be formulated as

$$y_l = \mathcal{D}(C(y_1, y_2, \ldots, y_{l-1})), \tag{1}$$

where y_l denotes the output feature map of the lth layer, $C(\cdot)$ refers to the concatenation of the output feature maps of previous layers. Motivated by [42] and [45], $\mathcal{D}(\cdot)$ is defined as a composite function of three connected operations: batch normalization (BN), a rectified linear unit (ReLU), and a dilated convolution. The dilated convolution operation can be formulated as

$$y_{l(i,j)} = \sum_u \sum_v x_{l(r\times u, r\times v)} \cdot k_{(i-u, j-v)}, \tag{2}$$

where r is the dilated rate, u and v are the coordinate offsets in k, k is the dilated convolution kernel, and $y_{l(i,j)}$ is the value of the lth layer's output feature map at (i, j).

The diverse dilated rates in different ConvBlocks generate a deep and dense pyramid hierarchical architecture that increases the scale of feature extraction with the increase of the receptive field, thus contributing to the search for the LV structure in multi-scale space. The dilated rates of five ConvBlock are 1, 1, 2, 4, and 8, respectively. The input feature maps of all levels are propagated to the ConvBlock after a convolution layer except levels 2 and 3, where an average pooling operation is inserted after the convolution layer. The channel numbers of the output feature maps of five abstraction levels are compressed to 72, 100, 114, 121, and 20 separately by the convolution layer. Moreover, the output feature map of every ConvBlock is also convolved to generate a highly abstract feature map (denoted as $f_1 \sim f_5$) before being sent to the skip connections for the feature fusion. With the help of the pyramid ConvBlocks, the semantic gap between the highly abstract feature maps ($f_1 \sim f_5$) and feature maps of the feature fusion network is greatly reduced compared with plain skip connections.

Behind the deep feature pyramid module, inspired by [43], a pyramid pooling module (as depicted in Fig. 5) is embedded to extract further global contextual information and reduce global contextual information loss in the largest receptive field. The pyramid pooling module adopts four parallel poolings and convolutions to represent the feature maps at different scales (1×1, 2×2, 3×3 and 6×6). Then, four represented feature maps are upsampled to the same scale and concatenated together, and next convolved by a 1×1 kernel to reduce the number of channels. The output feature map of the pyramid pooling module has 4 channels and is sent to the feature fusion network.

The deep feature pyramid module joint the pyramid pooling module endows DPSN with the superior feature extraction ability and the LV region detection capacity in multi-level and multi-scale space, further contributing to the capture of the global geometric characteristic of the LV and the establishment of uniform semantic features.

3.2. Multi-level and multi-scale features fusion

This part also consists of a multi-level pyramid hierarchical architecture. Multi-level and multi-scale holistic semantic features are fused to generate multiple predictions at different abstraction levels, helping to restore high resolution segmentation result with different semantic degrees. As shown in the middle of Fig. 3, the feature maps ($f_1 \sim f_5$) produced by the pyramid pooling module and by five ConvBlocks are concatenated hierarchically from level 5 to level 1, allowing DPSN to propagate contextual semantic information to higher resolution levels. An upsampling operation is required to match the scale of next level's feature map when the concatenation operation comes to level 1 and level 2. The bilinear interpolation with transposed convolution is utilized in the upsampling operation.

In addition, the concatenated feature maps of all five abstraction levels are upsampled to 256×256 (the same as the input image) to generate multiple predictions (denoted as $P_1 \sim P_5$). Next, $P_1 \sim P_5$ are fused to generate the final prediction P_0. During the training process, $P_0 \sim P_5$ are compared with the ground truth to produce multiple losses (denoted as $loss_0 \sim loss_5$) by the composite loss function described in the next section.

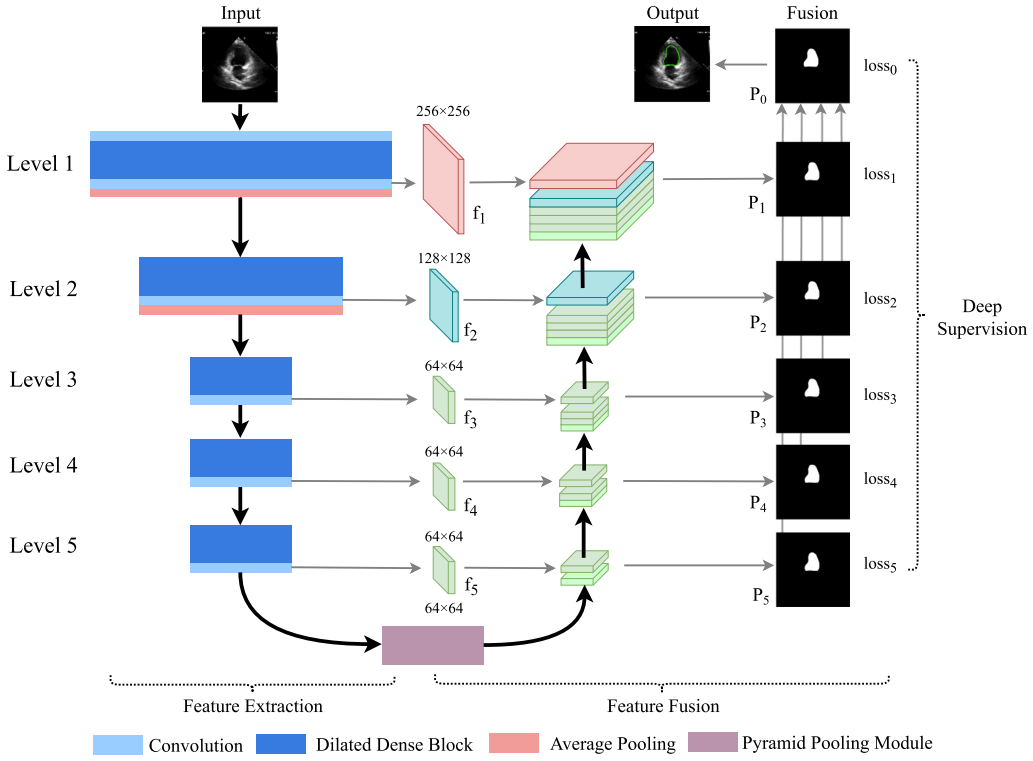

Fig. 3. Schematic view of DPSN. It consists of the deep feature pyramid module, the pyramid pooling module, and the deep supervision module. $f_1 \sim f_5$ are feature maps extracted from the ConvBlocks of five levels. $P_1 \sim P_5$ are multi-level predictions, and P_0 is the final prediction fused by $P_1 \sim P_5$.

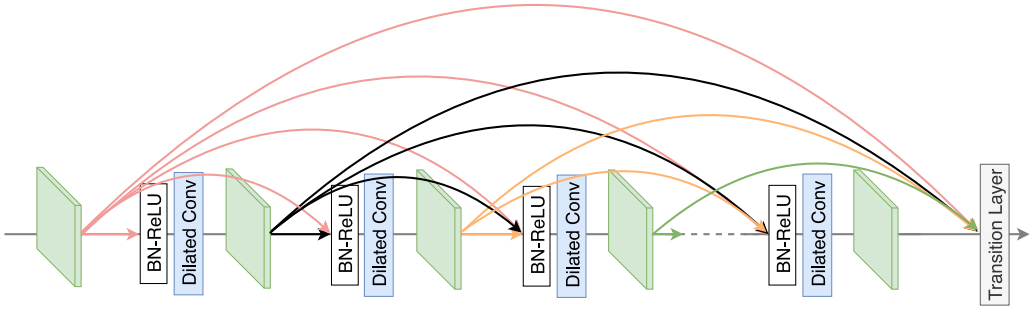

Fig. 4. Dilated dense convolution block. Dilated convolution layers displace the plain convolution layers. The transition layers between two joint ConvBlocks change feature map channels and sizes via convolution and pooling.

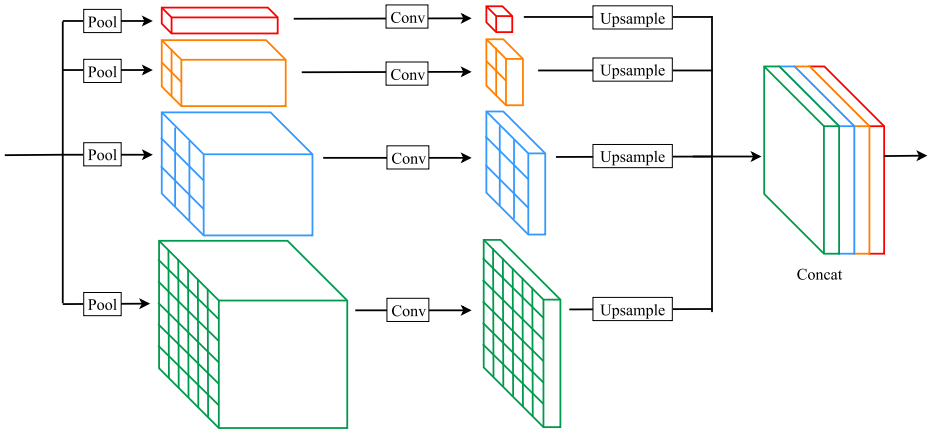

Fig. 5. Pyramid pooling module.

Table 1
Specifications of heterogeneous dataset from multiple centers and vendors.

Center	Vendor machine	Patients	Train	Test	SNR
The Third People's Hospital of Shenzhen	Philips EPIQ 7C	60	48	12	21.13 ± 3.76
Peking University First Hospital	GE VIVID E9	20	16	4	17.90 ± 4.81
The First Affiliated Hospital of Shenzhen University	Philips IE33	20	16	4	23.18 ± 2.49
–	Total	100	80	20	–

3.3. Hierarchical deep supervision

DPSN outputs full resolution feature maps at all abstraction levels with the help of modified skip pathways and the upsampling operation, and further generates multiple predictions ($P_0 \sim P_5$), which are amenable to develop into deep supervision [44]. The predictions of all levels are exposed to the ground truth to introduce the composite loss function to every feature fusion level. Multiple loss functions generate multi-level hierarchical deep supervision, which helps to enhance the gradient signal during the backpropagation. The loss function at each level is utilized to introduce newly computed gradients and reduce the difference between $P_i (i \in \{0, \ldots, 5\})$ and the ground truth, respectively. They directly propagate feedback to all convolution layers in the feature extraction network, thereby minimizing the gradient vanishing problem. Meanwhile, under the effect of hierarchical deep supervision, competition among multiple predictions and mutual regularization help to effectively alleviate the over-fitting problem.

Hierarchical deep supervision helps to minimize the gradient vanishing and over-fitting problems by directly building feedback from the ground truth to every convolution layer. This mechanism also boosts the hierarchical information flow and fits the latent hierarchical features in fine scales, which helps to constrain the LV boundaries and learn a better semantic representation, thereby enabling DPSN to obtain superior LV segmentation performance. All these factors contribute to achieving good generalization and robustness, and further facilitate the deep layer architecture training process.

Composite losses. A composite loss function $\mathcal{L}$ is proposed that contains three aspects for the comparison between multiple predictions and the ground truth: pixel-level similarity, overlapping degree and spatial Euclidean distance. During the training process, a pixel-wise softmax activation is applied to every level in the feature fusion network to generate multiple predicted probability maps.

With regard to the pixel-level similarity, the pixel-wise weighted binary cross-entropy loss ($\mathcal{L}_{WCE}$) [23] is utilized and formulated as

$$\mathcal{L}_{WCE}(G, P) = -\lambda \cdot w \cdot G \cdot log\ P - (1 - G) \cdot log(1 - P), \tag{3}$$

where G and P denote the flattened ground truth and the predicted probabilities, respectively. The weight w is a trade-off between the segmented territory and the other part, formulated as $w = \frac{1-P}{P}$; λ is an additional parameter chosen empirically during the training process and used to control the influence of w.

Regarding the overlapping degree, the generalized dice loss ($\mathcal{L}_{GD}$), a generalized dice index modified from the dice score coefficient [46], is used to evaluate the segmentation performance. It can be expressed as

$$\mathcal{L}_{GD}(G, P) = 1 - 2\frac{\alpha \cdot G \cdot P + \beta \cdot (1 - G) \cdot (1 - P)}{\alpha \cdot (G + P) + \beta \cdot (2 - G - P)}, \tag{4}$$

where α and β are two parameters used to provide invariance to different label set properties, formulated as $\alpha = G^{-2}$ and $\beta = P^{-2}$, respectively.

Table 2
Details of the multi-view dataset.

Dataset	A2C	A3C	A4C	Total	Patients
Train	2870	2891	2889	8650	80
Test	745	732	731	2208	20
Total	3615	3623	3620	10858	100

For the spatial Euclidean distance, the softplus function is utilized to modify the mean absolute error and obtain the modified mean absolute error loss ($\mathcal{L}_{MMAE}$), facilitating the optimization of this loss function:

$$\mathcal{L}_{MMAE}(G, P) = log\left(1 + e^{|G-P|}\right), \tag{5}$$

A combination of the above three loss functions is applied to the prediction at every abstraction level ($P_1 \sim P_5$) as well as to the final fusion prediction (P_0), which is described as

$$\mathcal{L} = \sum_{i=0}^{5} \lambda_i \cdot \Big(\lambda_{WCE} \cdot \mathcal{L}_{WCE}(G, P_i) + \lambda_{GD} \cdot \mathcal{L}_{GD}(G, P_i) + \lambda_{MMAE} \cdot \mathcal{L}_{MMAE}(G, P_i) \Big), \tag{6}$$

where λ_{WCE}, λ_{GD}, and λ_{MMAE} are the corresponding balance coefficients of $\mathcal{L}_{WCE}$, $\mathcal{L}_{GD}$, and $\mathcal{L}_{MMAE}$, respectively; λ_i represents the corresponding balance coefficients of $loss_i$ (as depicted in Fig. 3). All λs are chosen empirically during the training process.

In summary, as depicted in Fig. 3, DPSN surpasses the original U-Net from four aspects: (1) the pyramid ConvBlocks embedded in the feature extraction network helps to lessen the semantic gap and promote the flow of gradients and contextual information; (2) the dense and deep pyramid hierarchical architecture enables the model to combine multi-level and multi-scale features efficiently; (3) the pyramid pooling module bridges the feature extraction network and feature fusion network and reduces contextual information loss; and (4) the introduction of hierarchical deep supervision with composite losses to the feature fusion network enhances the gradient signal in the backpropagation process, which efficiently alleviates the over-fitting problem.

4. Experiments and results

4.1. Dataset and evaluation metrics

The segmentation performance of DPSN is validated on a multi-view heterogeneous echocardiographic video dataset, which was acquired from several different vendor machines at three hospitals (as depicted in Table 1). Each patient has three videos, namely, A2C, A3C, and A4C. Each video consists of several cardiac cycles; but only one cardiac cycle is extracted, and the endocardial borders were labeled by experts. Fig. 6 presents A2C, A3C, and A4C samples segmented by DPSN (red) and experts (green). Finally, 10858 labeled images were acquired, consisting of 300 sequences from 100 patients, as illustrated in Table 2. To the best of our knowledge, this is the largest labeled echocardiographic sequence dataset constructed to date.

Geometrical metrics. To evaluate the segmentation quality, the F-measure (F_β), the Dice metric, the Hausdorff distance

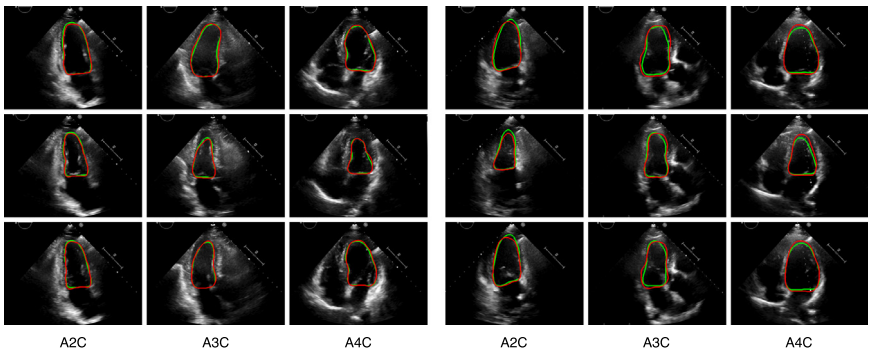

Fig. 6. Good (left part) and bad (right part) automatic segmentation examples (red) compared with manually traced contours labeled by experts (green). Top row: ED frames; Middle row: ES frames; Bottom row: frames in the middle of the cardiac cycle.

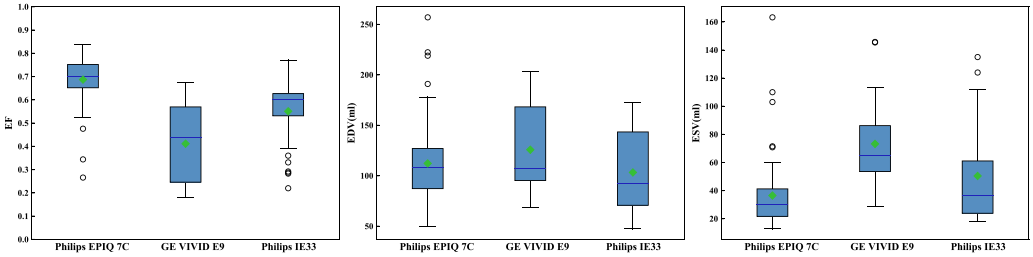

Fig. 7. Statistical analyses of clinical indices (EF, EDV, and ESV) across centers and vendors.

(HD[mm]), and the mean absolute distance (MAD[mm]) are adopted. Let P and G be the contour segmented by DPSN and by experts, respectively, and let Ω_P and Ω_G be the areas enclosed by contours P and G, respectively. F-measure evaluates the detection performance. It is depicted as $F_\beta(G,P) = \frac{(1+\beta^2)\cdot Precision\cdot Recall}{\beta^2\cdot Precision+Recall}$, where β is a parameter used to achieve a tradeoff between the precision and recall ($Precision = \frac{|\Omega_G\cap\Omega_P|}{|\Omega_P|}$, $Recall = \frac{|\Omega_G\cap\Omega_P|}{|\Omega_G|}$). Here, $\beta^2 = 0.3$ is set to give more weight to the precision than to the recall [47]. Dice indicates the mutual overlap between the ground truth and segmentation result. HD measures the largest distance between two boundaries. While MAD measures the average boundary distance between two boundaries. The entire image is used as the segmentation result to calculate MAD and HD if no segmented object is generated.

Clinical metric. In addition to evaluating the segmentation quality, the estimation of ejection fraction (EF), which is more clinically relevant, is performed. EF is calculated according to the standard guidelines [7].

Statistical analyses of heterogeneous data. The signal-to-noise ratio (SNR) of different centers and vendors is calculated. As shown in Table 1, distinct gaps exist among different centers and vendors. The SNRs of Philips EPIQ 7C and IE33 are much better than that of GE VIVID E9. EF, end-diastolic volume (EDV), and end-systolic volume (ESV) of different centers and vendors are also compared. As depicted in Fig. 7, the clinical indices of different centers and vendors show significant differences. These results reflect the gaps in heterogeneous data from different centers and vendors.

4.2. Implementation details

For the sake of computational efficiency, all echocardiographic images are downsampled to 256 × 256 before being sent to DPSN. In the training process, the Stochastic Gradient Descent (SGD) with a momentum of 0.9 is employed as the optimization algorithm to minimize the proposed composite loss function $\mathcal{L}$. The Xavier scheme [48] is selected to initialize the weights of parameters randomly. L_2 regularization with a weight decay coefficient of 0.0001 is utilized as a supplement to minimize the over-fitting problem. In addition, a polynomial decay strategy, expressed as $lr = lr_0 \times \left(1 - \frac{epoch}{30}\right)^{0.5}$, is utilized to reduce the learning rate, where $lr_0 = 0.002$ is the initial learning rate; 0.5 is the decay rate; and *epoch* is the current iteration number. The training process contains 50 epochs.

On the training set, ten-fold cross-validation is utilized to provide an unbiased estimation, while the testing set is an independent dataset (i.e., data unseen by the trained model) used for the final performance evaluation. All experiments are performed on a Linux workstation equipped with an Intel Xeon 2.10 GHz CPU and a 12 GB NVidia Titan XP GPU using the TensorFlow framework.

4.3. Ablation experiments

In order to illustrate the effectiveness of the proposed network architecture, a series of ablation experiments are conducted to analyze and quantify the importance and necessity of each proposed component in DPSN. Next, additional ablation experiments are performed to evaluate the robustness and generalization of DPSN on different views and across different centers and vendors.

Different DPSN variants. In this section, the performances of different ablated DPSN variants are compared. As shown in Fig. 8, each color represents a variant of DPSN under a specific configuration. From left to right, each color means respectively removing pyramid pooling module, dilated convolution (i.e., using the original DenseNet to replace the dilated dense blocks), deep supervision (i.e., employing only the composite loss function $\mathcal{L}$ on P_1 while neglecting multiple losses and multiple predictions fusion mechanism) from the original DPSN and successively reducing the levels of the original DPSN (i.e., reducing the depth of DPSN). The results show that the full DPSN achieves higher mean values of F_β and Dice and lower mean values of HD and

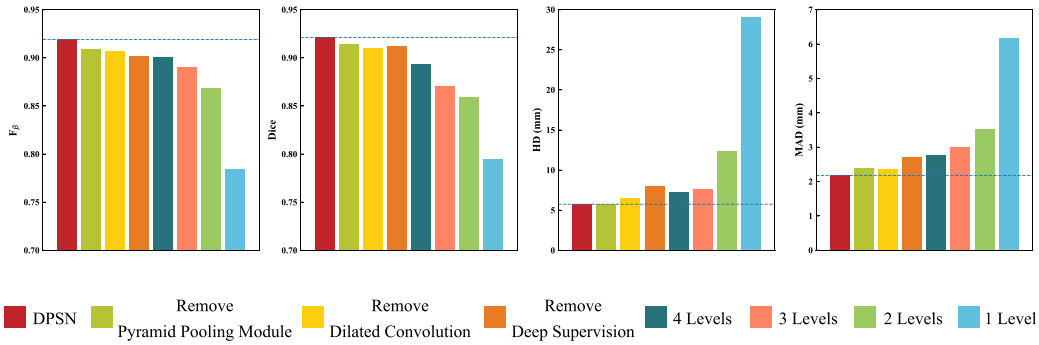

Fig. 8. Ablation experiments of different DPSN variants. Different colors represent DPSN variants with different configurations. The horizontal dashed lines indicate the performance of the full DPSN.

Table 3
The robustness and generalization of DPSN on different views. Sample means and standard deviations (denoted as Mean ± SD) of the segmentation by DPSN.

View	F_β	Dice	HD (mm)	MAD (mm)
A2C	0.936 ± 0.042	0.942 ± 0.029	4.87 ± 1.56	1.61 ± 0.71
A3C	0.927 ± 0.069	0.932 ± 0.037	5.27 ± 2.27	1.98 ± 1.03
A4C	0.931 ± 0.053	0.939 ± 0.032	4.99 ± 1.97	1.71 ± 0.83
A234C	0.919 ± 0.057	0.921 ± 0.046	5.75 ± 3.14	2.18 ± 1.22

MAD compared to its variants. Each individual component significantly improves the LV segmentation performance, especially when increasing the depth of DPSN.

Different loss function variants. To verify the efficiency of the proposed composite loss function $\mathcal{L}$, different variants of $\mathcal{L}$ are further evaluated. As illustrated in Fig. 9, different colors represent the loss function under different configurations: from left to right, respectively, are the proposed composite loss function $\mathcal{L}$, $\mathcal{L}_{GD}$ combined with $\mathcal{L}_{MMAE}$, $\mathcal{L}_{WCE}$ combined with $\mathcal{L}_{GD}$, and $\mathcal{L}_{WCE}$ combined with $\mathcal{L}_{MMAE}$. The composite loss function $\mathcal{L}$ achieves higher mean values of F_β and Dice and relatively lower mean values of HD and MAD compared to three other loss function configurations. These results indicate that every component of $\mathcal{L}$ is useful.

Evaluation of DPSN on different views. To certify the robustness and generalization of DPSN on different views, it is trained and tested on A2C, A3C, A4C, and A234C (i.e., all three views) sequence datasets separately. As shown in Table 3, first, DPSN is evaluated on three different views; the means and standard deviations of F_β, Dice, HD and MAD values indicate that DPSN performs well on A2C, A3C and A4C separately. When using all three views (A234C) at the same time, DPSN can still achieve high mean values of F_β (0.919 ± 0.057) and Dice (0.921 ± 0.046), low mean values of HD (5.75 ± 3.14 mm) and MAD (2.18 ± 1.22 mm), as well as significantly low standard deviations on all the metrics. The experimental results show the superior segmentation performance of DPSN on single-view echocardiographic sequences and its remarkable robustness and generalization on multi-view echocardiographic sequences.

Evaluation of DPSN across different centers and vendors. To understand the robustness and generalization of DPSN across different centers and vendors, it is trained on multi-view data from one center and vendor but tested with sequences from another center and vendor. As shown in Table 4, compared with testing on the same center and vendor, testing on another center and vendor yields comparable means and standard deviations of F_β, Dice, HD, and MAD. The performance of DPSN experiences only a moderate decline when applied to different centers and vendors. These experimental results demonstrate the robustness and generalization of DPSN across different centers and vendors.

4.4. Performance comparison

DPSN is compared with four methods: FCN [21], U-Net [32], iMD-FCN [29], and ACNN [28]. First, all the methods are compared on the ED and ES frames of the full heterogeneous dataset to illustrate that DPSN can cope with ED and ES frames segmentation, and outperform the other methods. Next, they are compared on multi-view echocardiographic sequences of the full heterogeneous dataset to further demonstrate that DPSN can achieve superior segmentation accuracy on echocardiographic sequences.

Comparison study on ED and ES frames. The comparison results are shown in Table 5. DPSN achieves the best segmentation performance on the ED and ES frames of the full heterogeneous dataset; it obtains the highest means and lowest standard deviations of F_β, indicating that it can capture the anatomical structure of the LV more robustly than the other methods. DPSN also achieves the highest mean values of Dice, the lowest mean values of HD and MAD, as well as substantially lower standard deviations on all the metrics. These results demonstrate that when addressing the ED and ES frames segmentation, DPSN surpasses the other methods by a large margin whether on the region coverage, contour accuracy, or distance error.

Comparison study on echocardiographic sequences. Table 6 presents the comparison results. DPSN still outperforms all the other methods on all the metrics, achieving the highest mean values of F_β (0.919 ± 0.057) and Dice (0.921 ± 0.046), the lowest mean values of HD (5.75 ± 3.14 mm) and MAD (2.18 ± 1.22 mm), and significantly lower standard deviations on all the metrics. These results strongly demonstrate that DPSN is able to achieve the best region coverage, the highest contour accuracy, and the minimum distance error when processing multi-view echocardiographic sequences.

4.5. Stability evaluation

Although each frame is processed independently without utilizing temporal information, the stability of the segmentation results is still an important aspect of the performance evaluation. In clinical studies, unstable performance is unacceptable for cardiac cycle evaluation. The performance stability reflects how well the segmentation results match between consecutive frames. Adequate experiments are conducted as follows to demonstrate that DPSN can achieve stable and coherent performance throughout the cardiac cycle, even without utilizing temporal information. Additionally, to certify the stability of DPSN in the face of noise and the advantages of leaving out temporal modeling, an interesting experiment is designed that compares DPSN with 3D FCN-LSTM [49] on the dataset under different noise levels.

Performance stability. The means of F_β, Dice, HD, and MAD at different frames of all sequences are computed to investigate

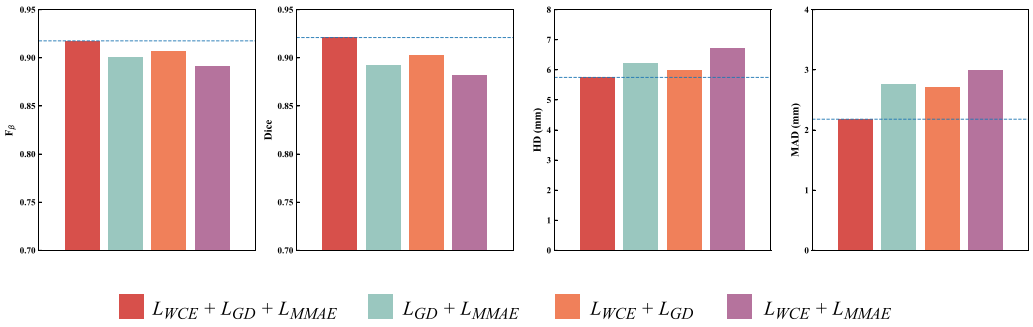

Fig. 9. Loss function ablation experiments. Different colors represent the loss function variants with different configurations. The horizontal dashed lines indicate the performance of the composite loss function $\mathcal{L}$.

Table 4
The robustness and generalization of DPSN across different centers and vendors. Sample means and standard deviations (denoted as Mean ± SD) of the segmentation by DPSN.

Train	Test	F_β	Dice	HD (mm)	MAD (mm)
Philips EPIQ 7C	Philips EPIQ 7C	0.926 ± 0.050	0.931 ± 0.040	5.33 ± 2.98	1.97 ± 0.99
Philips EPIQ 7C	Philips IE33	0.921 ± 0.060	0.927 ± 0.044	5.35 ± 3.01	2.09 ± 1.20
Philips EPIQ 7C	GE VIVID E9	0.913 ± 0.071	0.919 ± 0.054	6.79 ± 4.82	3.02 ± 2.07
Philips IE33	Philips IE33	0.930 ± 0.047	0.933 ± 0.036	5.29 ± 2.75	1.80 ± 1.01
Philips IE33	Philips EPIQ 7C	0.920 ± 0.058	0.925 ± 0.047	5.52 ± 3.37	2.11 ± 1.33
Philips IE33	GE VIVID E9	0.917 ± 0.064	0.920 ± 0.051	6.22 ± 4.31	2.96 ± 1.89
GE VIVID E9	GE VIVID E9	0.923 ± 0.062	0.929 ± 0.042	5.57 ± 3.04	2.01 ± 1.13
GE VIVID E9	Philips EPIQ 7C	0.911 ± 0.073	0.918 ± 0.060	7.07 ± 4.98	3.24 ± 1.97
GE VIVID E9	Philips IE33	0.908 ± 0.080	0.915 ± 0.066	6.89 ± 4.90	3.56 ± 2.19

Table 5
Sample means and standard deviations (denoted as Mean ± SD) of the segmentation quality measures for multi-view ED and ES frames across multi-vendor and multi-center.

Method	F_β		Dice		HD (mm)		MAD (mm)	
	ED	ES	ED	ES	ED	ES	ED	ES
Proposed	0.933 ± 0.042	0.917 ± 0.075	0.945 ± 0.025	0.925 ± 0.049	5.31 ± 2.69	5.70 ± 3.05	1.61 ± 0.52	1.93 ± 0.72
FCN	0.883 ± 0.062	0.872 ± 0.071	0.894 ± 0.040	0.881 ± 0.054	7.41 ± 5.67	8.56 ± 5.69	2.76 ± 1.23	3.36 ± 1.92
U-Net	0.919 ± 0.047	0.890 ± 0.068	0.934 ± 0.031	0.905 ± 0.050	6.96 ± 3.77	7.95 ± 4.51	1.98 ± 0.91	2.26 ± 1.01
iMD-FCN	0.911 ± 0.041	0.891 ± 0.074	0.928 ± 0.033	0.904 ± 0.052	7.13 ± 4.16	8.06 ± 4.87	1.99 ± 1.02	2.41 ± 1.08
ACNN	0.920 ± 0.037	0.908 ± 0.069	0.939 ± 0.023	0.913 ± 0.046	6.61 ± 2.96	7.57 ± 2.66	1.89 ± 0.71	2.20 ± 0.92

Table 6
Sample means and standard deviations (denoted as Mean ± SD) of the segmentation quality measures for multi-view echocardiographic sequences across multi-vendor and multi-center.

Method	F_β	Dice	HD (mm)	MAD (mm)
Proposed	0.919 ± 0.057	0.921 ± 0.046	5.75 ± 3.14	2.18 ± 1.22
FCN	0.839 ± 0.086	0.852 ± 0.079	10.34 ± 9.81	4.21 ± 2.69
U-Net	0.869 ± 0.077	0.883 ± 0.068	8.94 ± 6.87	3.72 ± 1.87
iMD-FCN	0.862 ± 0.081	0.879 ± 0.070	9.04 ± 7.01	3.99 ± 2.01
ACNN	0.882 ± 0.073	0.893 ± 0.061	7.70 ± 6.58	3.40 ± 1.57

the volatility of each metric. All the echocardiographic sequences are limited to 30 frames for the stability evaluation. As depicted in Fig. 10, DPSN achieves stable mean values on all four metrics in the cardiac cycle, as well as moderate fluctuating standard deviations. The experimental results demonstrate that DPSN is able to segment each frame independently with high accuracy and achieve stable and coherent performance in the sequence, even without utilizing temporal information. The "by-product" (stable and coherent performance) is actually due to the robust, accurate LV anatomical structure detection and segmentation accomplished by DPSN.

Comparison study on image sequences with different noise levels. Ten noisy datasets with different noise levels are generated by adding noise randomly to the original dataset and gradually increasing the noise quantity. Finally, eleven noisy datasets are obtained, including the original dataset. As shown in Fig. 11, from beginning to end, DPSN always achieves higher mean values of F_β and Dice, as well as lower mean values of HD and MAD compared with 3D FCN-LSTM. Moreover, as the noise quantity increases, the performance of DPSN remains relatively stable, while the performance of 3D FCN-LSTM fluctuates dramatically. Based on the above experimental results, it can be concluded that DPSN is more robust to noise disturbances. DPSN processes all frames independently instead of utilizing temporal information, which minimizes the information loss and avoids the accumulation of temporal error in noisy conditions. In temporal modeling, LSTM is sensitive to noise disturbances and not robust, and even small disturbances can cause large changes in the subsequent results. In addition, the loop in LSTM amplifies the noise and disturbs the representation learning, causing frequent information loss.

4.6. Clinical quantification

EF is an important cardiac functional index and predictor used for prognosis. The LV volume and EF are computed according to the standard guidelines [7]. The correlation coefficient and Bland–Altman agreement are adopted to achieve a comprehensive assessment.Fig. 12 presents the correlation analysis and Bland–Altman analysis of the comparison of the EF measurements computed from automatic segmentations by DPSN (EF_a) and manual segmentations by experts (EF_m). DPSN achieves

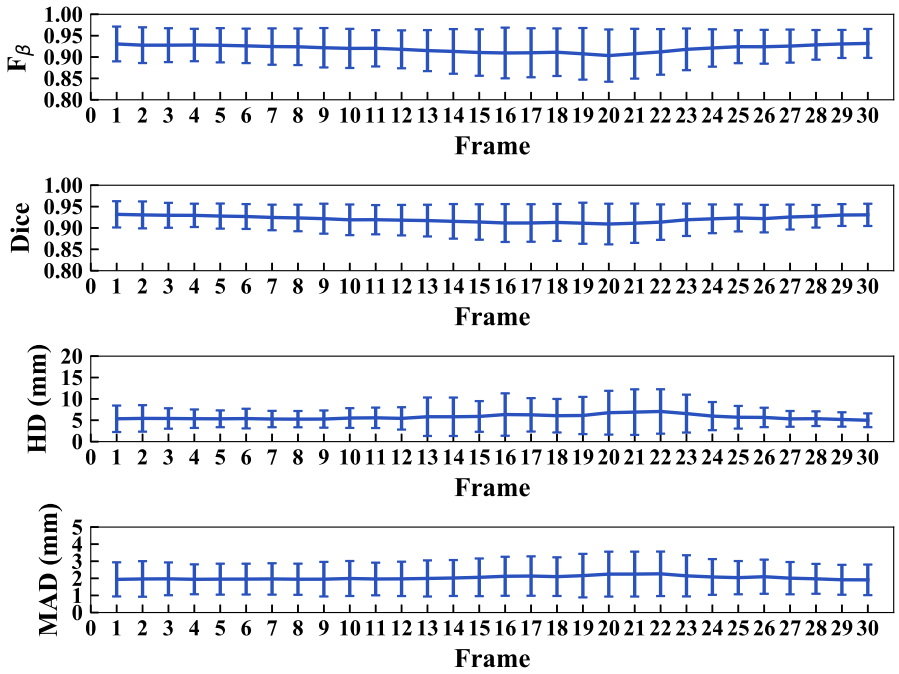

Fig. 10. Means of F_β, Dice, HD, and MAD at different frames of the cardiac cycle.

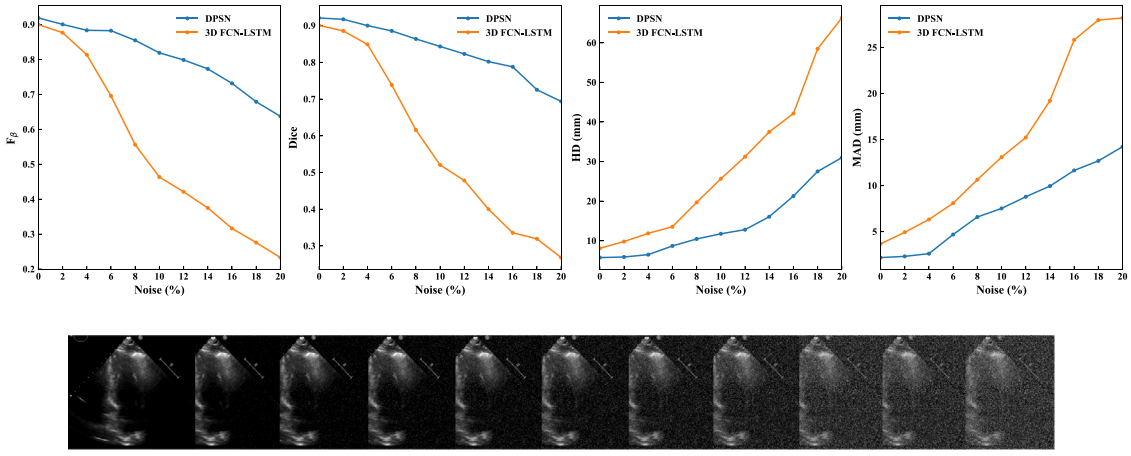

Fig. 11. Comparison between DPSN and 3D FCN-LSTM on image sequences with different noise levels.

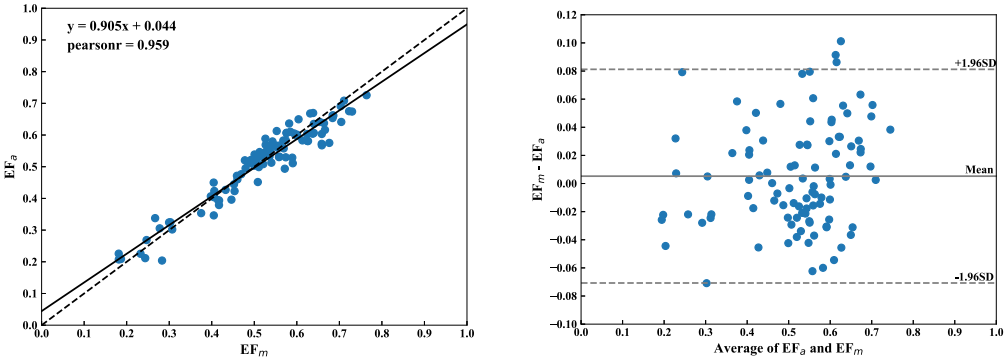

Fig. 12. Correlation analysis (left) and Bland–Altman analysis (right) showing the agreement between the EF measurements calculated from automatic segmentations by DPSN (EF_a) and manual segmentations by experts (EF_m).

high correlation (0.959) and agreement (0.0052 ± 0.0759) on EF compared with experts. The mean difference between EF_a and EF_m is 0.0052. The 95% limits of agreement (mean difference ± standard deviation) are [0.081, −0.071]. 96% of the measurements are located in the ±1.96 standard deviation in the Bland–Altman plot. These results reveal the clinical potential of DPSN.

5. Conclusion

This paper presents a deep pyramid and deep supervision network that achieves a unified model for multi-view echocardiographic sequences interpretation without utilizing temporal information. DPSN extracts and fuses multi-level and multi-scale holistic semantic features efficiently, obtaining prominent generalization and robustness to accommodate heterogeneous data.

Instead of temporal modeling, DPSN processes each frame in the sequence independently, which eliminates temporal error accumulation and frequent information loss. This has the inherent advantages of being robust to noise disturbances and reducing the computational complexity.

The geometrical and clinical evaluation experiments demonstrate that DPSN achieves: superior segmentation accuracy on heterogeneous data as a result of the prominent generalization and robustness; stable and coherent performance as a "by-product" of the independent processing, rather than as the result of an explicitly imposed and expensive constraint.

Because DPSN is a supervised learning based method, it requires large amounts of labeled data. However, building large amounts of labeled data takes a lot of time and effort. In future work, a semi-supervised learning based method will be introduced to reduce the reliance on large amounts of labeled data.

Declaration of competing interest

No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.asoc.2019.106049.

CRediT authorship contribution statement

Ming Li: Conceptualization, Writing - original draft, Methodology, Formal analysis. **Shizhou Dong:** Software. **Zhifan Gao:** Methodology, Supervision. **Cheng Feng:** Data curation, Validation. **Huahua Xiong:** Data curation, Validation. **Wei Zheng:** Supervision, Funding acquisition. **Dhanjoo Ghista:** Writing - review & editing. **Heye Zhang:** Supervision, Funding acquisition. **Victor Hugo C. de Albuquerque:** Writing - review & editing.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. U1801265 and No. 61771464), the Guangdong Science and Technology Department (No. 2018A050506031 and No. 2019B010110001), the Guangdong Natural Science Funds for Distinguished Young Scholar (No. 2019B151502031), the Fundamental Research Funds for the Central Universities, and the Brazilian National Council for Scientific and Technological Development (CNPq) via Grants Nos. 304315/2017-6, and 430274/2018-1.

References

[1] J.A. Noble, D. Boukerroui, Ultrasound image segmentation: a survey, IEEE Trans. Med. Imaging 25 (8) (2006) 987–1010.
[2] X. Huang, D.P. Dione, C.B. Compas, X. Papademetris, B.A. Lin, A. Bregasi, A.J. Sinusas, L.H. Staib, J.S. Duncan, Contour tracking in echocardiographic sequences via sparse representation and dictionary learning, Med. Image Anal. 18 (2) (2014) 253–271.
[3] T. Plappert, M.G.S.J. Sutton, The Echocardiographers' Guide, CRC Press, 2006.
[4] S. Osher, N. Paragios, Geometric Level Set Methods in Imaging, Vision, and Graphics, Springer Science & Business Media, 2003.
[5] A. Madani, R. Arnaout, M. Mofrad, R. Arnaout, Fast and accurate view classification of echocardiograms using deep learning, NPJ Digit. Med. 1 (1) (2018) 6.
[6] E.J. Topol, High-performance medicine: the convergence of human and artificial intelligence, Nature Med. 25 (1) (2019) 44–56.
[7] R.M. Lang, L.P. Badano, V. Mor-Avi, J. Afilalo, A. Armstrong, L. Ernande, F.A. Flachskampf, E. Foster, S.A. Goldstein, T. Kuznetsova, et al., Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the american society of echocardiography and the european association of Cardiovascular imaging, Eur. Heart J.-Cardiovasc. Imaging 16 (3) (2015) 233–271.
[8] C.M. Otto, Textbook of Clinical Echocardiography E-Book, Elsevier Health Sciences, 2013.
[9] P. Salen, R. O'Connor, P. Sierzenski, B. Passarello, D. Pancu, S. Melanson, S. Arcona, J. Reed, M. Heller, Can cardiac sonography and capnography be used independently and in combination to predict resuscitation outcomes?, Acad. Emerg. Med. 8 (6) (2001) 610–615.
[10] C.L. Moore, G.A. Rose, V.S. Tayal, D.M. Sullivan, J.A. Arrowood, J.A. Kline, Determination of left ventricular function by emergency physician echocardiography of hypotensive patients, Acad. Emerg. Med. 9 (3) (2002) 186–193.
[11] J.G. Bosch, S.C. Mitchell, B.P. Lelieveldt, F. Nijland, O. Kamp, M. Sonka, J.H. Reiber, Automatic segmentation of echocardiographic sequences by active appearance motion models, IEEE Trans. Med. Imaging 21 (11) (2002) 1374–1383.
[12] F. Milletari, A. Rothberg, J. Jia, M. Sofka, Integrating statistical prior knowledge into convolutional neural networks, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2017, pp. 161–168.
[13] S. Leclerc, T. Grenier, F. Espinosa, O. Bernard, A fully automatic and multi-structural segmentation of the left ventricle and the myocardium on highly heterogeneous 2d echocardiographic data, in: 2017 IEEE International Ultrasonics Symposium (IUS), IEEE, 2017, pp. 1–4.
[14] D. Barbosa, D. Friboulet, J. D'hooge, O. Bernard, Fast tracking of the left ventricle using global anatomical affine optical flow and local recursive block matching, MIDAS J. 10 (2014).
[15] J. Pedrosa, S. Queiros, O. Bernard, J. Engvall, T. Edvardsen, E. Nagel, J. D'Hooge, Fast and fully automatic left ventricular segmentation and tracking in echocardiography using shape-based b-spline explicit active surfaces, IEEE Trans. Med. Imaging 36 (2017) 2287–2296.
[16] S. Queirós, J.L. Vilaça, P. Morais, J.C. Fonseca, J. D'hooge, D. Barbosa, Fast left ventricle tracking using localized anatomical affine optical flow, Int. J. Numer. Methods Biomed. Eng. 33 (11) (2017) e2871.
[17] W. Xue, G. Brahm, S. Pandey, S. Leung, S. Li, Full left ventricle quantification via deep multitask relationships learning, Med. Image Anal. 43 (2018) 54–65.
[18] C. Xu, L. Xu, G. Brahm, H. Zhang, S. Li, Mutgan: Simultaneous segmentation and quantification of myocardial infarction without contrast agents via joint adversarial learning, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2018, pp. 525–534.
[19] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.
[20] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: International Conference on Learning Representations, 2014.
[21] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.
[22] H. Noh, S. Hong, B. Han, Learning deconvolution network for semantic segmentation, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1520–1528.
[23] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 234–241.
[24] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, J. Dean, A guide to deep learning in healthcare, Nature Med. 25 (January) (2019).
[25] L.K. Tan, Y.M. Liew, E. Lim, R.A. McLaughlin, Convolutional neural network regression for short-axis left ventricle segmentation in cardiac cine mr sequences, Med. Image Anal. 39 (2017) 78–86.
[26] D.M. Vigneault, W. Xie, C.Y. Ho, D.A. Bluemke, J.A. Noble, Ω-Net (omega-net): Fully automatic, multi-view cardiac mr detection, orientation, and segmentation with deep neural networks, Med. Image Anal. 48 (2018) 95–106.
[27] G. Carneiro, J.C. Nascimento, A. Freitas, 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 (3) (2012) 968–982.

[28] O. Oktay, E. Ferrante, K. Kamnitsas, M. Heinrich, W. Bai, J. Caballero, S.A. Cook, A. de Marvao, T. Dawes, D.P. O'Regan, et al., Anatomically constrained neural networks (acnns): application to cardiac image enhancement and segmentation, IEEE Trans. Med. Imaging 37 (2) (2018) 384–395.
[29] H. Chen, Y. Zheng, J.-H. Park, P.-A. Heng, S.K. Zhou, Iterative multi-domain regularized deep learning for anatomical structure detection and segmentation from ultrasound images, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2016, pp. 487–495.
[30] R.P. Poudel, P. Lamata, G. Montana, Recurrent fully convolutional neural networks for multi-slice mri cardiac segmentation, in: Reconstruction, Segmentation, and Analysis of Medical Images, Springer, 2016, pp. 83–94.
[31] M.-M. Rohé, M. Datar, T. Heimann, M. Sermesant, X. Pennec, Svf-net: Learning deformable image registration using shape matching, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2017, pp. 266–274.
[32] J. Zhang, S. Gajjala, P. Agrawal, G.H. Tison, L.A. Hallock, L. Beussink-Nelson, M.H. Lassen, E. Fan, M.A. Aras, C. Jordan, et al., Fully automated echocardiogram interpretation in clinical practice: feasibility and diagnostic accuracy, Circulation 138 (16) (2018) 1623–1635.
[33] Q. Tao, W. Yan, Y. Wang, E.H. Paiman, D.P. Shamonin, P. Garg, S. Plein, L. Huang, L. Xia, M. Sramko, et al., Deep learning–based method for fully automatic quantification of left ventricle function from cine mr images: a multivendor, multicenter study, Radiology 290 (1) (2018) 81–88.
[34] Z. Gao, Y. Li, Y. Sun, J. Yang, H. Xiong, H. Zhang, X. Liu, W. Wu, D. Liang, S. Li, Motion tracking of the carotid artery wall from ultrasound image sequences: a nonlinear state-space approach, IEEE Trans. Med. Imaging 37 (1) (2018) 273–283.
[35] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, L. Van Gool, One-shot video object segmentation, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2017, pp. 5320–5329.
[36] K.-K. Maninis, S. Caelles, Y. Chen, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, L. Van Gool, Video object segmentation without temporal information, IEEE Trans. Pattern Anal. Mach. Intell. 41 (6) (2018) 1515–1530.
[37] B. Kong, Y. Zhan, M. Shin, T. Denny, S. Zhang, Recognizing end-diastole and end-systole frames via deep temporal regression network, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2016, pp. 264–272.
[38] C. Xu, L. Xu, Z. Gao, S. Zhao, H. Zhang, Y. Zhang, X. Du, S. Zhao, D. Ghista, H. Liu, et al., Direct delineation of myocardial infarction without contrast agents using a joint motion feature learning architecture, Med. Image Anal. 50 (2018) 82–94.
[39] J. Bosch, J.H. Reiber, Two-dimensional echocardiographic digital image processing and approaches to endocardial edge detection, in: The Practice of Clinical Echocardiography, Saunders, Philadelphia, 2002, pp. 141–158.
[40] T. Laurent, J. von Brecht, A recurrent neural network without chaos, in: International Conference on Learning Representations, 2016.
[41] M. Chen, Minimalrnn: Toward more interpretable and trainable recurrent neural networks, in: Conference and Workshop on Neural Information Processing Systems, 2017.
[42] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks., in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, Vol. 1, 2017, p. 3.
[43] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2881–2890.
[44] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, Z. Tu, Deeply-supervised nets, in: Artificial Intelligence and Statistics, 2015, pp. 562–570.
[45] K. He, X. Zhang, S. Ren, J. Sun, Identity mappings in deep residual networks, in: European Conference on Computer Vision, Springer, 2016, pp. 630–645.
[46] T. Liu, J. Sun, N.-N. Zheng, X. Tang, H.-Y. Shum, Learning to detect a salient object, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2007, pp. 1–8.
[47] T. Wang, A. Borji, L. Zhang, P. Zhang, H. Lu, A stagewise refinement model for detecting salient objects in images, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4019–4028.
[48] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feed-forward neural networks, in: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, 2010, pp. 249–256.
[49] X. Yang, L. Yu, S. Li, H. Wen, D. Luo, C. Bian, J. Qin, D. Ni, P.-A. Heng, Towards automated semantic segmentation in prenatal volumetric ultrasound, IEEE Trans. Med. Imaging 38 (1) (2019) 180–193.