IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 23, 2019, accepted January 21, 2020, date of publication February 3, 2020, date of current version February 17, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2971383

AIDAN: An Attention-Guided Dual-Path Network for Pediatric Echocardiography Segmentation

YUJIN HU[1], BEI XIA[2], MUYI MAO[2], ZELONG JIN[2], JIE DU[1], LIBAO GUO[1], ALEJANDRO F. FRANGI[3,4], (Fellow, IEEE), BAIYING LEI[1], (Senior Member, IEEE), AND TIANFU WANG[1]

[1]National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, Guangdong Key Laboratory for Biomedical Measurements and Ultrasound Imaging, Health Science Center, School of Biomedical Engineering, Shenzhen University, Shenzhen 518060, China
[2]Department of Ultrasound, Shenzhen Children's Hospital, Shenzhen 518050, China
[3]Centre for Computational Imaging & Simulation Technologies in Biomedicine (CISTIB), School of Computing, University of Leeds, Leeds LS2 9JT, U.K.
[4]School of Medicine, University of Leeds, Leeds LS2 9JT, U.K.

Corresponding authors: Bei Xia (xiabeimd@qq.com), Baiying Lei (leiby@szu.edu.cn), and Tianfu Wang (tfwang@szu.edu.cn)

This work was supported in part by the National Key R&D Program of China (2016YFC0104700), National Natural Science Foundation of China under Grant 81571758, in part by the Shenzhen Key Basic Research Project under Grant JCYJ20180507184647636, Grant JCYJ20170818142347251, and Grant JCYJ20170818094109846, in part by the AFF was funded by the Royal Academy of Engineering under Grant INSILEX CiET1819/19, in part by the Chinese Academy of Sciences (PIFI Program), and in part by the European Commission under Grant H2020-SC1-PM-16-2017-777119.

ABSTRACT Accurate segmentation of pediatric echocardiography images is essential for a wide range of diagnostic and pre-interventional planning, but remains challenging (e.g., low signal to noise ratio and internal variability in heart appearance). To address these problems, in this paper, we propose a novel Cardiac Attention-guided Dual-path Network (i.e., AIDAN). AIDAN comprises a convolutional block attention module (CBAM) attached to a spatial (i.e., SPA) and context paths (i.e., CPA), which can guide the network and learn the most discriminative features. The spatial path captures low-level spatial features, and the context path is designed to exploit high-level context. Finally, features learned from the two paths are fused efficiently using a specially designed feature fusion module (FFM), and these are used to predict the final segmentation map. We experiment on a self-collected dataset of 127 pediatric echocardiography cases which are videos containing at least a complete cardiac cycle, and obtain a Dice coefficient of 0.951 and 0.914, in the left ventricle and atrium segments, respectively. AIDAN outperforms other state-of-the-art methods and has great potential for pediatric echocardiography images analysis.

INDEX TERMS Convolutional block attention module, dual-path network, feature fusion module, pediatric echocardiography segmentation.

I. INTRODUCTION

Congenital heart disease (CHD) is a type of birth defect with abnormal heart and vessels structures, related to environmental and genetic factors of the fetus or the pregnant women[1]. There are 1.5∼2 million children born with CHD according to the World Health Organization [2]. Krasuski [1] reported that the incidence of CHD in America is nearly 1%, and that in China is 1.42% [3]. Echocardiography is the primary examination method for CHD diagnosis. It is non-invasive, low-cost and suitable for real-time imaging. Accurate segmentation of cardiac anatomy in echocardiography images is an essential step for a wide range of analysis and diagnosis, such as measuring the ejection fraction [4, 5]. However, this task highly relies on manual segmentation, which brings a heavy burden to sonographers.

In early clinical work, cardiac segmentation was achieved by manual delineation of anatomical boundaries of the heart, which is time-consuming, subjective and error prone. Many researchers attempted to address this challenge in various imaging modalities [6]–[12]. These researchers are mainly focused on magnetic resonance imaging in adults with less attention given to pediatric cardiac ultrasound analysis [13]. Both adult and pediatric cardiac segmentation share similar challenges in poor image quality of ultrasound, but pediatric heart is more variable and complex in terms of

The associate editor coordinating the review of this manuscript and approving it for publication was Vishal Srivastava.

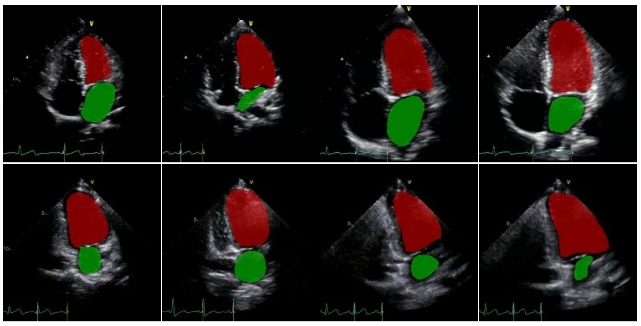


FIGURE 1. Examples of echocardiography images for 4CH (the first row) and 2CH (the last row) view.

morphology and appearance than adult heart. Thus, pediatric cardiac segmentation is particularly challenging. Though some automatic or semi-automatic methods were proposed to solve adult cardiac segmentation [9], [14], [15] and pediatric cardiac segmentation [16]–[18], there are still remaining challenges. Fig. 1 shows the typical 4 chamber (4CH) (first row) and 2 chamber (2CH) (second row) view echocardiography images with segmentation masks. Accurate pixel-wise segmentation of left ventricle (LV) and left atrium (LA) in echocardiography images suffer from these challenges: 1) low signal to noise ratio, varying amounts of speckle noise, and presence of shadowing in the ultrasound images; 2) the influence of different scan styles, image protocols, and different devices, e.g., the first two columns and the last two columns come from different devices; 3) intra-variability of children heart appearances.

To address the above challenges, deep learning has been widely applied due to its impressive performance across many tasks [19]–[27]. Deep learning methods also gained popularity across medical image analysis [9], [28]–[34]. Owing to great success of convolutional neural networks (CNNs), many researchers adopted them in medical image analysis [11], [19], [29]–[32], [34]–[37]. However, some of these CNNs [6], [10], [16], [32] do not consider low-level and high-level features simultaneously. Some also rely on complex extra post-processing procedures [21], [24] (e.g., conditional random field [38], [39], CRF). Hence, these methods could not be trained in an end-to-end fashion. Besides, even though U-Net based methods [10], [12], [40] consider fusing both low-level spatial and high-level semantic context features, but are in an inefficient way. To overcome these limitations, BiSeNet is proposed. BiSeNet uses a spatial path and a context path to capture low-level and high-level features, respectively, and then fuses the learned features for better predictions [41]. To further refine BiSeNet, a convolutional block attention module (CBAM), is proposed in [42] for efficient feature weighting, and guide the network in learning the most discriminative features.

Inspired by previous works, a novel end-to-end framework coined AIDAN is proposed in this paper and applied to pediatric echocardiography segmentation. AIDAN can extract both low-level and high-level features by a dual-path network. It also can automatically extract the most discriminative features, and then fuse them effectively, which finally accurately segments both LV and LA effectively. This new end-to-end neural network for pediatric echocardiography segmentation makes these contributions:

1) A spatial path and a context path are used for capturing the low-level spatial features and the high-level context features, respectively;

2) The attention model, CBAM, which focuses on both ''what'' and ''where'' to look, is adopted in both the spatial path (i.e., SPA) and the context path (i.e., CPA) to guide the network to extract the most discriminative features;

3) A feature fusion module (FFM) is used to fuse features learned from the spatial path and context path at different scales efficiently.

The remainder of this paper is organized as follows. Section II introduces the related work. The proposed method is described in Section III. The experiments and comparison results are illustrated in Section IV. Our discussions are given in Section V. Finally, our conclusions are presented in Section VI.

II. RELATED WORKS

A. SEGMENTATION METHODS WITHOUT ATTENTION MECHANISMS

Fully convolutional neural network (FCN) [24] and U-Net [40] are the first two neural networks attempting semantic segmentation in a fully convolutional manner in computer vision and biomedical image analysis, respectively. Both FCN and U-Net use encoder-decoder architectures, where the encoder is designed for feature extraction, and the decoder is responsible for up sampling to reconstruct a semantic segmentation mask with the same size of input images. Bai *et al.* [8] proposed to use FCN for cardiac magnetic resonance (MR) image analysis. The main difference between FCN and U-Net lies in that U-Net concatenates the features from the encoder to the decoder at the same layer to account for spatial information while FCN does not. The major drawback of FCN lies in losing spatial information. DeepLab improves on FCN by using dilated convolutions [21]. PSPNet further improves FCN by using pyramid pooling to learn multi-scale features [26]. In biomedical image analysis, incorporating medical or anatomical priors to neural networks has shown to improve CNN performance. For example, Veni *et al.* [10] proposed to segment 4CH from echocardiography images with U-Nets [40] combined with anatomical shape priors. Duan *et al.* [43] also considered anatomical shape constraints. They proposed to feed 2.5D representation of input cardiac MR cine into a FCN and refine the segmentation with explicitly enforcing a shape constraint [43]. Oktay *et al.* proposed another anatomically constrained FCN [7], which incorporates the anatomical priors for learning, that is, learning a latent representation through TL-Network [44]. Though incorporating anatomical priors into FCN improves the interpretability, it requires expert knowledge and annotated data that might not always be available at scale. Zheng *et al.* [6] proposed to segment cardiac MR

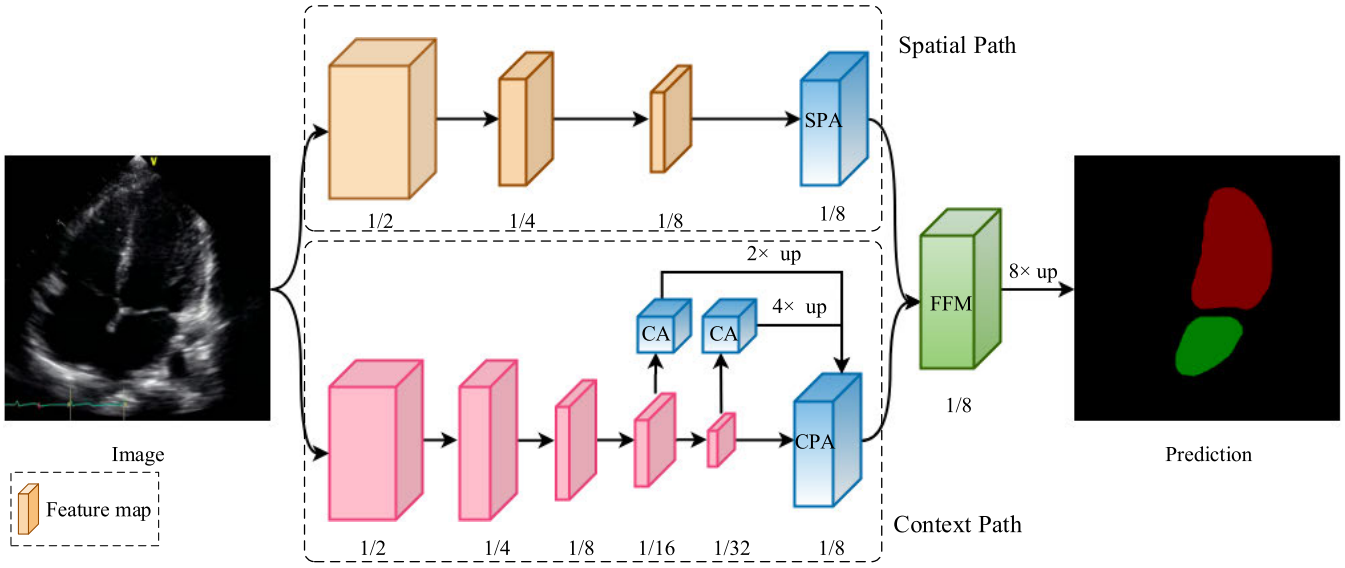


FIGURE 2. Overview of the proposed AIDAN. It consists of a spatial path, a context path and an FFM. The spatial path captures low-level spatial features, while the con-text path captures high-level context features. A CBAM module is attached to the spatial path and the context path.

cines using U-Net with spatial propagation. Given that the correlation between slices is low except for adjacent slices, instead of feeding entire MR volume to the network, they argue that feeding the current slice and its previous segmentation prediction to a 2D U-Net can explicitly maintain the 3D consistency. Therefore, this method captures 3D spatial context features using a 2D U-Net, which is computational efficient. OmegaNet is also proposed to segment cardiac MR cines by Vigneault *et al.* [12]. OmegaNet first learns to predict an initial segmentation result using an hourglass network [12], and then the parameters needed for transforming the input image to a canonical orientation are predicted using these learned features. The final segmentation is learned from the transformed image using stacked hourglass networks [45]. Though OmegaNet provides greater flexibility by exploiting intermediate segmentations, it requires additional annotations (e.g. parameters for the affine transformation from the input image to a canonical orientation).

B. SEGMENTATION METHODS WITH ATTENTION MECHANISMS

Attention mechanisms have been widely used in segmentation tasks [23], [25], [27], [42]. The attention mechanism is realized in a channel-wise and spatial-wise fashion. The squeeze and excitation block adaptively recalibrates channel-wise feature responses by explicitly modelling interdependencies between channels, and comprises a channel attention mechanism [27]. Meanwhile, non-local blocks proposed in [25] computes the response at a position as a weighted sum of the features at all positions, and introduces spatial attention. Later on, based on non-local block, Fu *et al.* [23] proposed to append two types of attention modules (e.g., position attention module and channel attention module) on top of traditional dilated FCN to improve the performance. BiSeNet proposed an attention refinement module, which is similar to an squeeze and excitation block, but employs global average pooling to capture global context [46]. Our previous work [47] has successfully applied BiSeNet for pediatric echocardiography image segmentation. Instead of using additional layers, Zhang *et al.* [28] proposed to modify the original residual block [20] by using feature maps learned in a high layer as attention map for a low layer.

In this paper, we propose a novel dual-path convolutional neural network, AIDAN. AIDAN takes advantage of BiSeNet's dual-path design to capture both low-level spatial and high-level context features. AIDAN further extends BiSeNet by incorporating CBAM to guide the network to learn the most discriminative features.

III. METHOD

As shown in Fig. 2, the proposed AIDAN consists of a spatial path, a context path, and an FFM. Besides, a CBAM is added to the spatial path and the context path for better feature extraction, denoted as SPA and CPA in Fig. 2, respectively. The loss function used in the proposed model is also introduced in the following subsection III-E. The network details are presented in the following sections.

A. SPATIAL PATH

Rich spatial information and large receptive fields are considered crucial in segmentation. Dilated convolutions are usually used [21], [26] to preserve rich spatial information in semantic segmentation. Some approaches try to capture sufficient receptive fields with pyramid pooling, atrous spatial pyramid pooling or "large kernel" [26], [48]. These methods show that the spatial information and the receptive field are both crucial to semantic segmentation. Based on this observation, the spatial path preserves the spatial size of the input image while encoding rich spatial information [46]. It consists of only three blocks, each block includes a 3×3 convolutional layer with stride $= 2$ followed by batch normalization and

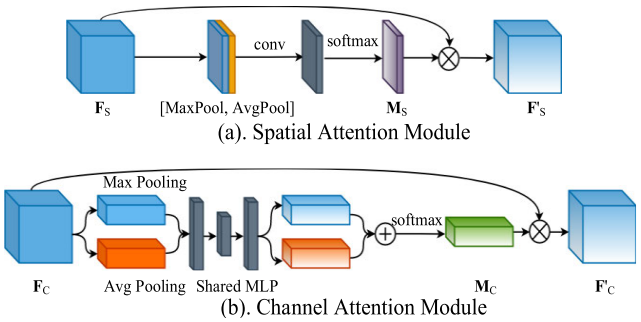


FIGURE 3. Overview of CBAM. (a). This sub-module utilizes the inter-spatial relationship of features and focuses on "where" an informative part is. (b). This sub-module utilizes the inter-channel relationship of features and focuses on "what" is meaningful.

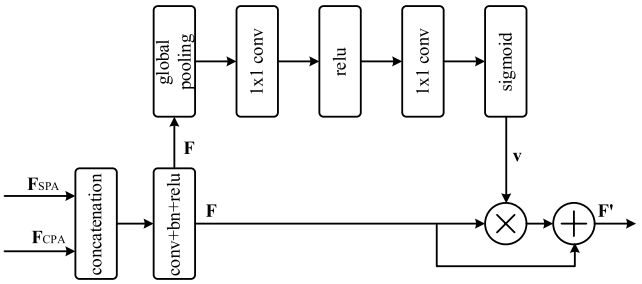


FIGURE 4. Feature fusion module, features learned from the spatial path $\mathbf{F_{SPA}}$ and features learned from the context path $\mathbf{F_{CPA}}$ are first concatenated, and then balanced scales after batch normalization are used to produce F, and a weight vector v is computed to reweigh the feature, finally the refined feature $\mathbf{F}' = \mathbf{F} * \mathbf{v} + \mathbf{F}$.

ReLU activated layer. Therefore, the outputs feature maps of this path are of 1/8 size of the original input image and it encodes rich low-level spatial information at a low cost of computation.

B. CONTEXT PATH

While the spatial path encodes rich spatial information, the context path captures a sufficient receptive field and high-level semantic context features. To enlarge the receptive field, previous approaches take advantage of the pyramid pooling module, atrous spatial pyramid pooling or large convolutional kernel. However, the methods mentioned above all suffer from heavy computation demanding and memory consuming. The context path is used in this method to get a large receptive field with efficient computation [46]. The residual network ResNet50 [20] is utilized as our backbone network. To further refine the features learned from the residual context path, a channel attention module (denoted as CA in Fig. 2) is used to refine features learned at the last two down sampling stages. CA is a submodule of CBAM.

C. CONVOLUTIONAL BLOCK ATTENTION MODULE

The CBAM consists of a spatial attention sub-module and a channel attention sub-module. The spatial attention sub-module utilizes the inter-spatial relationship of features. As shown in Fig. 3a, given an input intermediate feature map $\mathbf{F}_s$, we firstly aggregate the spatial information by max pooling and average pooling along the channel axis. We concatenate them to generate an efficient feature descriptor. Then a convolutional layer is applied to the concatenated feature descriptor to compute a 2D spatial attention map $\mathbf{M_s}$. Finally, the refined feature $\mathbf{F}'_s$ is produced by multiplying the input feature $\mathbf{F}_s$ with the spatial attention map $\mathbf{M_s}$. Thus, the spatial attention sub-module exploits the inter-spatial relationship and focuses on ''where'' an informative part is.

As illustrated in Fig. 3b, the channel attention sub-module is designed for utilizing the inter-channel relationship of features. As each channel of a feature map is a feature detector, given an input image or feature map, channel attention focuses on ''what'' is meaningful. First, average pooling and max pooling are applied to the input feature $\mathbf{F}_c$ to generate two spatial context descriptors. Then both descriptors are fed into a shared multi-layer perceptron (MLP) with only one hidden layer. To reduce parameter overhead, the hidden layer unit is set to C/r, where C is the number of input feature maps, r is the reduction ratio. After the shared MLP is applied to the two feature descriptors, their output feature vectors are merged by element-wise summation, followed by a softmax layer to produce a weighted vector $\mathbf{M}_c$. Finally, the input feature $\mathbf{F}_c$ is multiplied with the weighed vector $\mathbf{M}_c$ to produce the refined feature $\mathbf{F}'_c$.

These two sub-modules in CBAM can be arranged in spatial-first, channel-first or in parallel. The influence of different CBAM module arrangements is illustrated in Section IV-C. Figure 3 shows a CBAM module added to the spatial path (SPA), and another added to the context path (CPA). Therefore, SPA is used for refining the features learned from the spatial path, and CPA is used for refining the features learned from the context path.

D. FEATURE FUSION MODULE

Obviously, the features learned from the spatial path and the context path are different in the level of feature representation. Therefore, simply summing up or concatenating these features is not a good fusion choice. Therefore, a feature fusion module proposed in [46] is explored in our network before the final prediction fusing features learned from the spatial and context paths. Specifically, as shown in Fig. 4, the features learned from the spatial path ($\mathbf{F}_{SPA}$) and the context path ($\mathbf{F}_{CPA}$) are first concatenated, then the batch normalization (BN) is utilized to balance the scales of the features to produce $\mathbf{F}$. The learned feature $\mathbf{F}$ is then pooled to a feature vector and a weight vector $\mathbf{v}$ is computed. Finally, the final refined feature $\mathbf{F}'$ is computed by $\mathbf{F}' = \mathbf{F} * \mathbf{v} + \mathbf{F}$.

E. LOSS FUNCTION

In this paper, the standard cross entropy loss is used for training the proposed network. Similar to the original BiSeNet [41], we also utilize the auxiliary loss function to supervise the training of the context path. Specifically, besides the principal loss function to supervise the output of the whole network, two specific auxiliary loss functions are also added to supervise the output of the context path at the

TABLE 1. 4CH view segmentation performance of AIDAN with different modules.

FFM	CA	SPA	CPA	Accuracy	Precision		Sensitivity		Specificity		Dice	
					LV	LA	LV	LA	LV	LA	LV	LA
				0.979 ±0.016	**0.960 ±0.050**	0.913 ±0.103	0.875 ±0.084	0.879 ±0.089	0.996 ±0.008	0.997 ±0.004	0.912 ±0.054	0.889 ±0.074
✓				0.978 ±0.017	0.919 ±0.045	0.912 ±0.110	0.928 ±0.094	0.887 ±0.081	**0.996 ±0.006**	0.997 ±0.004	0.922 ±0.062	0.893 ±0.072
✓	✓			0.983 ±0.010	0.926 ±0.048	0.868 ±0.101	0.949 ±0.046	0.928 ±0.099	0.983 ±0.006	0.983 ±0.005	0.936 ±0.031	0.889 ±0.084
✓	✓	✓		0.987 ±0.008	0.950 ±0.048	**0.919 ±0.086**	0.945 ±0.069	0.924 ±0.064	0.995 ±0.005	**0.998 ±0.003**	0.945 ±0.042	0.917 ±0.052
✓	✓	✓	✓	**0.987 ±0.007**	0.952 ±0.046	0.914 ±0.087	**0.950 ±0.050**	**0.929 ±0.060**	0.995 ±0.005	**0.998 ±0.003**	**0.949 ±0.031**	**0.918 ±0.051**

last two down sample stages. To summarize, the joint loss is:

$$L(\mathbf{X};\mathbf{W}) = L_p(\mathbf{X};\mathbf{W}) + \alpha_1 L_1(\mathbf{X_1};\mathbf{W}) + \alpha_2 L_2(\mathbf{X_2};\mathbf{W}) \quad (1)$$

where L_p is the principal loss, $\mathbf{X}$ is the final prediction of the whole network, $\mathbf{W}$ is the learnable parameters of the network, L_1 and L_2 are the auxiliary loss for the last two down sample stage of the context path, respectively. $\mathbf{X}_1$ and $\mathbf{X}_2$ are the output features from the last two down sample stage of the context path, respectively. The weight of the principal loss and auxiliary loss is balanced by α_1 and α_2. In this paper, $\alpha_1 = \alpha_2 = 1$ is set. The joint loss helps optimize the model more comfortable and easier.

IV. EXPERIMENTS AND RESULTS

A. DATASET

To verify the effectiveness of the proposed method, a dataset collected from Shenzhen Children Hospital is used for experimenting. All echocardiography images are collected from GE Vivid E8 and E9 (GE Healthcare, Horten, Norway) ultrasound machine. Our dataset consists of 127 video sequence of 4CH view. The age of the pediatric patient ranges from 0 to 10 years, the video frame rate is at least 24 fps, and each video sequence contains at least one complete cardiac cycle. We select 100 4CH videos randomly for training, and the remaining 27 videos are used for testing. To further verify the generalization of the proposed method, 12 2CH video sequences are used for testing only. Videos are further converted to images frame by frame. All images are manually segmented by two independent sonographers and additionally confirmed by a third experienced sonographer. Owing to poor images, a few frames might not be annotated, and these are removed from the dataset. Ultimately, we have 3654 4CH images for training, 831 4CH images and 503 2CH images for testing. In this paper, accuracy, precision (a.k.a. positive predictive value (PPV), recall (a.k.a. true positive rate (TPR), and sensitivity), specificity (a.k.a. true negative rate (TNR)), and Dice coefficient are used for estimating the segmentation performance.

B. IMPLEMENTATION PROTOCOL AND DATA AUGMENTATION

Our experiments are conducted on a computer workstation with Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10GHz, 4 GPU NVIDIA Titan Xp, and 64G of RAM, using PyTorch [49] and Horovo [50] for distributed training. During the training phase, the network parameters are updated using Adam optimizer. A poly learning rate strategy [21] is applied when training the network: $\eta = \eta_0 \left(1 - \frac{n}{N}\right)^{\beta}$, where $\eta_0 = 0.0001$ is the initial learning rate, n is the current epoch number, N is the total epochs, and $\beta = 0.9$. The pre-trained ResNet50 network from PyTorch is the backbone network of the context path to save time for training the network. All other components in the network are randomly initialized with PyTorch's default configuration.

All images are properly center cropped and resized to 512×512 to remove meaningless background or subject information. All images are normalized before fed to the neural networks. To avoid overfitting, common data augmentation methods, including random rotation $[-25, 25]$ degree, random flip, random crop, speckle noise, and salt & pepper noise of probability 0.01 are used during the training phase.

C. RESULTS AND ANALYSIS

We evaluate the performance of the proposed method through these three experiments: 1) Performance of the proposed method with different modules (FFM, CA, SPA, and CPA), 2) Performance of the proposed method with different CBAM module arrangements (spatial-first, channel-first or in parallel), 3) Performance comparison with state-of-the-art methods.

1) IMPACT OF DIFFERENT MODULES

To understand the influence of each module in our network and their contributions to the overall performance, we use AIDAN with no modules as the baseline: that is, a naive dual-path segmentation network with different network depths. Results are shown in the first row in Tables 1 and 2. Table 1 shows that for 4CH segmentation, FFM, CA, SPA, and CPA modules all improve the performance by 0.010, 0.014, 0.009, and 0.004 points in terms of Dice coefficient in LV segmentation, respectively. As for LA, it shows no improvement by adding CA, but FFM, SPA and CPA improve the performance by 0.004, 0.014 and 0.001 points in terms of Dice coefficient, respectively.

To explore the network generalization properties, we evaluate the optimized AIDAN with different modules

TABLE 2. 2CH view segmentation performance of AIDAN with different modules.

FFM	CA	SPA	CPA	Accuracy	Precision		Sensitivity		Specificity		Dice	
					LV	LA	LV	LA	LV	LA	LV	LA
				0.975	**0.906**	0.936	0.960	0.787	**0.993**	0.998	0.872	0.837
				±0.014	**±0.082**	±0.131	±0.142	±0.196	**±0.007**	±0.003	±0.102	±0.166
✓				**0.977**	0.829	0.912	0.915	0.691	0.992	**0.999**	0.862	0.754
				±0.011	±0.081	±0.086	±0.110	±0.128	±0.007	**±0.003**	±0.075	±0.096
✓	✓			0.967	0.791	0.918	0.933	0.674	0.981	0.998	0.851	0.729
				±0.013	±0.102	±0.166	±0.062	±0.302	±0.011	±0.003	±0.065	±0.0262
✓	✓	✓		0.976	0.833	**0.941**	**0.960**	**0.820**	0.985	0.998	0.886	**0.850**
				±0.013	±0.113	**±0.082**	**±0.043**	**±0.223**	±0.013	±0.002	±0.069	**±0.179**
✓	✓	✓	✓	0.976	0.839	0.932	0.958	0.807	0.985	0.998	**0.889**	0.834
				±0.012	±0.110	±0.119	±0.048	±0.248	±0.012	±0.024	**±0.068**	±0.211

TABLE 3. 4CH view segmentation performance of AIDAN with different CBAM module arrangement.

Arrangement	Accuracy	Precision		Sensitivity		Specificity		Dice	
		LV	LA	LV	LA	LV	LA	LV	LA
Channel-first	**0.987±0.007**	**0.952±0.046**	0.914±0.087	**0.950±0.050**	**0.929±0.060**	**0.995±0.005**	**0.998±0.003**	0.949±0.031	0.918±0.051
Spatial-first	0.986±0.008	0.952±0.047	0.914±0.084	0.950±0.067	0.919±0.065	**0.995±0.005**	**0.998±0.003**	**0.952±0.039**	0.914±0.050
Parallel	0.987±0.008	0.950±0.048	**0.923±0.085**	0.943±0.056	0.922±0.059	**0.995±0.005**	**0.998±0.003**	0.950±0.033	**0.923±0.049**

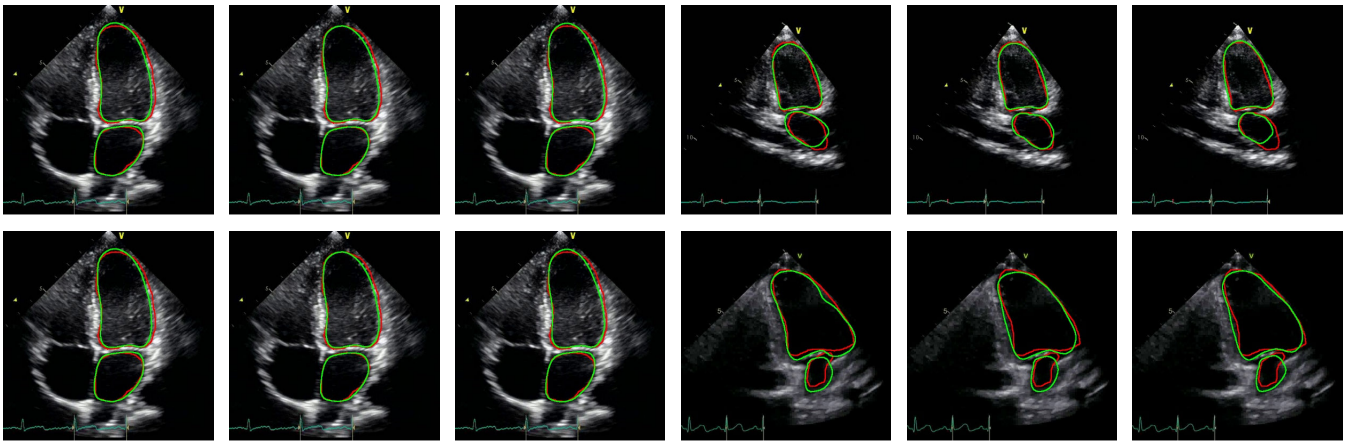


FIGURE 5. Segmentation results on AIDAN with different CBAM module arrangements, manual segmentation contours in red, and prediction results are in green. The first three columns are 4CH view images, while the last three columns are 2CH view images.

switched in/out. As shown in Table 2, AIDAN with all modules on in their best configuration performs the best in LV segmentation (best Dice coefficient), while AIDAN without CPA performs the best in LA segmentation. Given we only use 4CH images for training, and 2CH images for testing, our segmentation performance is promising.

2) IMPACT OF DIFFERENT CBAM MODULE ARRANGEMENTS

To show the influence of different CBAM module arrangements, we experiment spatial-first, channel-first, or both submodules in parallel. Table 3 shows that the parallel arrangement performs best for LV segmentation (Dice coefficient), while the spatial-first arrangement performs the best in LA segmentation. Sample segmentation results are shown in Fig. 5.

As for experiments on 2CH view images, the channel-first arrangement performs the best in the LV segmentation in terms of Dice coefficient, while the spatial-first arrangement performs the best in the LA segmentation.

As illustrated in Section III-C, context attention sub-module in CBAM focuses on "what" to look, while spatial attention sub-module focuses on "where". Therefore, we could conclude that, in LV segmentation, focusing on "where" is more important than focusing on "what", while focusing on "what" and "where" are almost equally important in LA segmentation. To get very good performance in both LV and LA segmentation, we recommend using the spatial-first arrangement in CBAM.

As shown in Table 4, the spatial-first arrangement performs the best in LA segmentation. Given that morphological variability in LA is richer and more complex, LA segmentation is more challenging than LV segmentation. The spatial-first arrangement shows great generalization ability from 4CH view images to 2CH images, this is another reason we recommend to use the spatial-first arrangement of CBAM module.

3) COMPARISON WITH THE STATE-OF-THE-ART METHODS

We benchmark our approach against several state-of-the-art methods, FCN [24], U-Net [40], DeepLab [21], PSPNet [26]

TABLE 4. 2CH View segmentation performance of AIDAN with different CBAM module arrangement.

Arrangement	Accuracy	Precision		Sensitivity		Specificity		Dice	
		LV	LA	LV	LA	LV	LA	LV	LA
Channel-first	**0.976±0.012**	0.839±0.110	0.932±0.119	**0.958±0.048**	0.807±0.248	0.985±0.012	0.998±0.024	**0.889±0.068**	0.834±0.211
Spatial-first	0.975±0.012	**0.842±0.112**	**0.943±0.091**	0.953±0.053	0.804±0.222	**0.986±0.012**	0.998±0.002	0.842±0.069	**0.943±0.165**
Parallel	**0.976±0.012**	0.837±0.106	0.938±0.116	0.957±0.052	**0.807±0.205**	0.985±0.012	**0.999±0.002**	0.837±0.066	0.938±0.169

TABLE 5. 4CH view segmentation performance comparison with state-of-the-art methods.

Network	Accuracy	Precision		Sensitivity		Specificity		Dice	
		LV	LA	LV	LA	LV	LA	LV	LA
FCN	0.975±0.020	0.949±0.052	0.841±0.152	0.884±0.079	0.893±0.073	0.995±0.007	0.993±0.010	0.912±0.050	0.857±0.099
U-Net	0.972±0.027	0.946±0.089	0.879±0.125	0.829±0.142	0.893±0.085	0.995±0.011	0.995±0.007	0.876±0.113	0.878±0.087
DeepLab	0.976±0.020	0.952±0.045	0.901±0.118	0.855±0.106	0.876±0.010	0.996±0.005	0.997±0.004	0.897±0.068	0.882±0.091
PSPNet	0.969±0.030	**0.965±0.049**	**0.930±0.095**	0.788±0.185	0.824±0.161	**0.998±0.003**	**0.998±0.003**	0.853±0.149	0.861±0.126
BiSeNet	0.984±0.009	0.937±0.068	0.918±0.087	0.938±0.043	0.898±0.080	0.994±0.008	**0.998±0.003**	0.935±0.039	0.903±0.056
AIDAN (channel-first)	**0.987±0.007**	0.952±0.046	0.914±0.087	**0.950±0.050**	**0.929±0.060**	0.995±0.005	**0.998±0.003**	0.949±0.031	0.918±0.051
AIDAN (spatial-first)	0.986±0.008	0.952±0.047	0.914±0.084	0.950±0.067	0.919±0.065	0.995±0.005	**0.998±0.003**	**0.952±0.039**	0.914±0.050
AIDAN (parallel)	0.987±0.008	0.950±0.048	0.923±0.085	0.943±0.056	0.922±0.059	0.995±0.005	**0.998±0.003**	0.950±0.033	**0.923±0.049**

TABLE 6. 2CH view segmentation performance comparison with state-of-the-art methods.

Network	Accuracy	Precision		Sensitivity		Specificity		Dice	
		LV	LA	LV	LA	LV	LA	LV	LA
FCN	0.972±0.017	0.898±0.075	0.922±0.113	0.827±0.203	0.751±0.215	0.993±0.006	0.998±0.004	0.840±0.164	0.800±0.168
U-Net	0.974±0.012	0.907±0.074	0.935±0.107	0.844±0.133	0.773±0.195	0.994±0.005	0.998±0.004	0.865±0.084	0.824±0.147
DeepLab	0.975±0.011	0.905±0.075	0.949±0.075	**0.963±0.099**	0.768±0.174	0.993±0.006	**0.999±0.002**	0.878±0.059	0.832±0.125
PSPNet	0.969±0.018	**0.927±0.064**	**0.959±0.087**	0.797±0.174	0.673±0.231	**0.995±0.004**	**0.999±0.002**	0.842±0.125	0.762±0.188
BiSeNet	0.969±0.012	0.837±0.112	0.918±0.130	0.848±0.145	**0.808±0.167**	0.987±0.010	0.998±0.002	0.829±0.095	0.848±0.146
AIDAN (channel-first)	**0.976±0.012**	0.839±0.110	0.932±0.119	0.958±0.048	0.807±0.248	0.985±0.012	0.998±0.024	**0.889±0.068**	0.834±0.211
AIDAN (spatial-first)	0.975±0.012	0.842±0.112	0.943±0.091	0.953±0.053	0.804±0.222	0.986±0.012	0.998±0.002	0.842±0.069	**0.943±0.165**
AIDAN (parallel)	**0.976±0.012**	0.837±0.106	0.938±0.116	0.957±0.052	0.807±0.205	0.985±0.012	**0.999±0.002**	0.837±0.066	0.938±0.169

and BiSeNet [46]. All these networks use ResNet-50 [20] so comparison can be done fairly and without bias. As shown in Table 5, AIDAN with channel-first arrangement, CBAM outperforms state-of-the-art methods in Dice coefficient in LV segmentation. AIDAN with parallel CBAM module arrangement performs the best among all the methods in Dice coefficient in LA segmentation.

The performance for all methods on 2CH view images are summarized in Table 6. We observe AIDAN with spatial-first arrangement in CBAM outperforms state-of-the-art methods in LV segmentation in Dice coefficient, and AIDAN with parallel arrangement performs the best Dice coefficient in LA segmentation.

Fig. 6a and Fig. 6b illustrate segmentation results averaging the Dice coefficient across all 27 subjects from test set. AIDAN outperforms state-of-the-art methods in both LV and LA segmentation on 4CH view images. Fig. 6c and Fig. 6d further illustrate AIDAN also performs comparably to state-of-the-art methods in LV and LA segmentation on unseen 2CH view images. This also highlights to the generalization ability of AIDAN. Sample segmentation results of 4CH and 2CH view images are shown in Fig. 7 and Fig. 8, respectively.

The reason for the promising segmentation performance of AIDAN is four-fold: 1) The spatial path captures low-level spatial features; 2) The context paths captures high-level semantic context features; 3) features learned from the dual paths are further refined by CBAM module to capture the most discriminative features; 4) The FFM fuses the features learned from the dual paths effectively.

V. DISCUSSION

To improve the segmentation of echocardiography image, an end-to-end AIDAN, is proposed in this paper. AIDAN effectively extracts features from both low-level and high-level paths. To demonstrate the effectiveness of the proposed AIDAN for LV and LA segmentation, we extract the feature maps and visualize them in Figs. 9 and 10. Specifically, we visualize the feature maps of CBAM module attached to both the spatial path and the context path. Our method can restore the original input image effectively. We analyze the effectiveness of our method in these aspects.

First, AIDAN is based on BiSeNet, which can capture low-level spatial features and high-level semantic features from the spatial path and the context path, respectively.

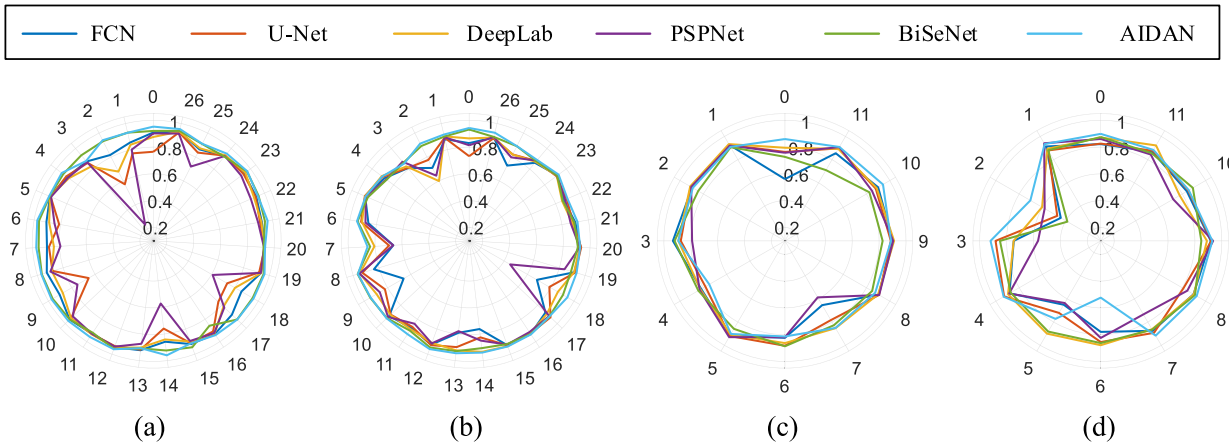


FIGURE 6. Dice value of all subjects on 4CH view and 2CH view of different methods, note that AIDAN with parallel arrangement of CBAM. (a) Dice value of 27 subjects on 4CH view of LV, (b) Dice value of 27 subjects on 4CH view of LA, (c) Dice value of 12 subjects on 2CH view of LV, (d) Dice value of 12 subjects on 2CH view of LA.

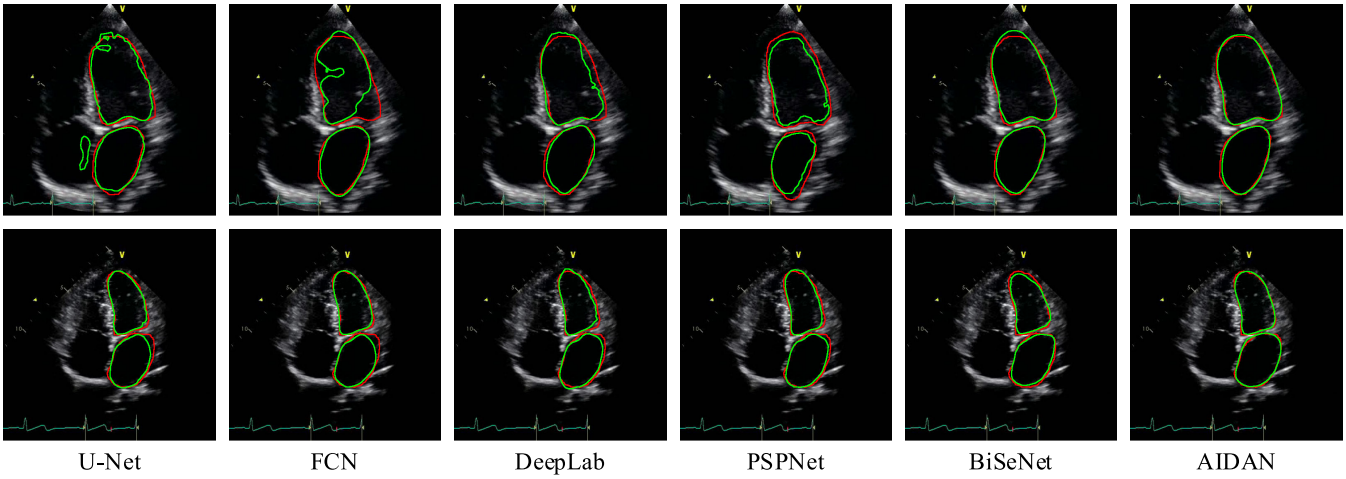


FIGURE 7. Segmentation result of 4CH view images on different networks, manual segmentation contours in red, prediction in green.

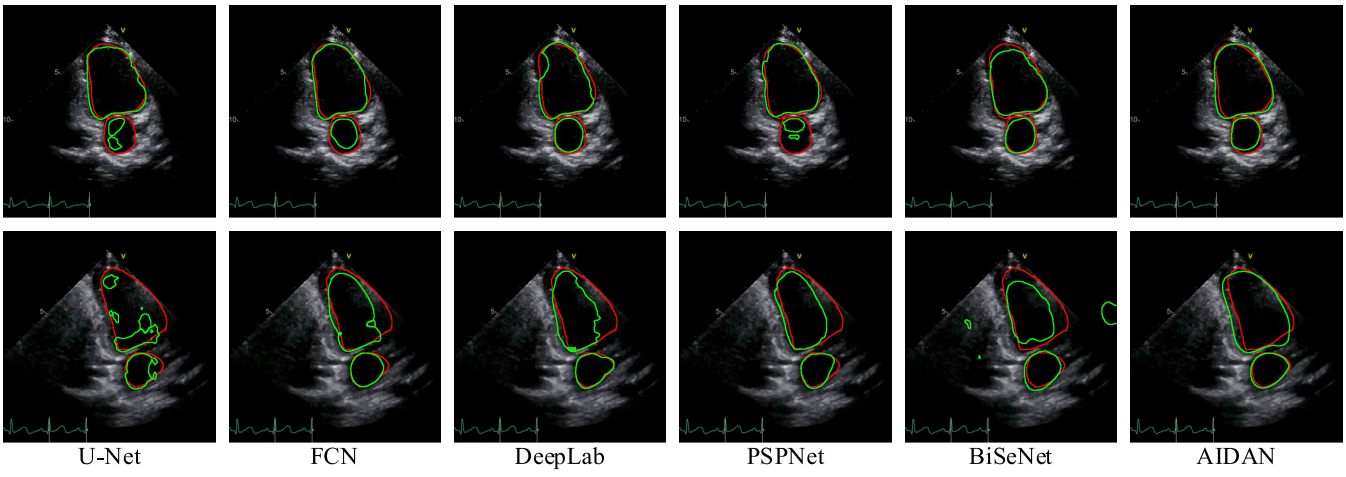
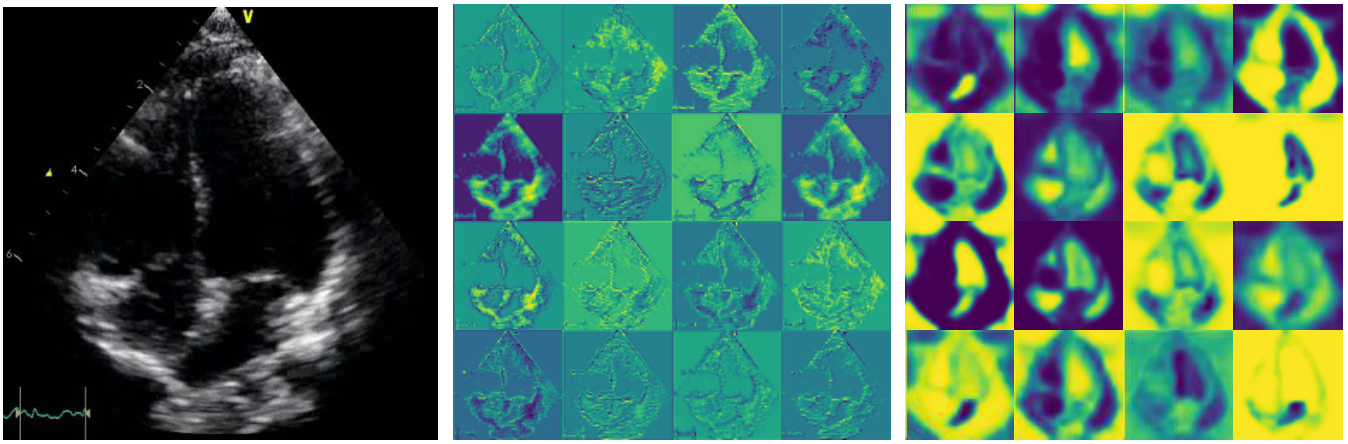


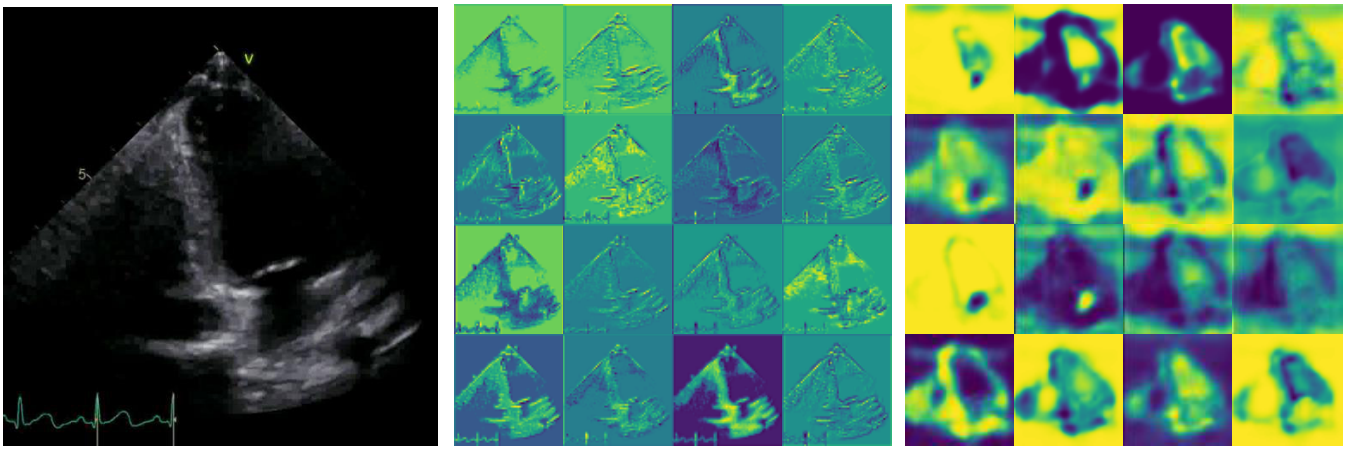
FIGURE 8. Segmentation result of 2CH view images on different networks, manual segmentation contours in red, prediction in green.

Second, instead of using skip connection to concatenate the feature learned at the encoding phase to the decoding phase, we use an CBAM attention module to refine the feature learned at the encoding phase. It can be regarded as an incomplete U-Net component. Finally, the features learned from the spatial path and the context path are fused using an FFM. Hence, the proposed AIDAN shows superiority over other related state-of-the-art methods.



(a) (b) (c)

FIGURE 9. Visualization of CBAM block with 4CH view image. (a) input image, (b) feature maps obtained by CBAM (SPA) attached to the spatial path, (c) feature maps obtained by CBAM (CPA) attached to the context path.



(a) (b) (c)

FIGURE 10. Visualization of CBAM block with 2CH view image. (a) input image, (b) feature maps obtained by CBAM (SPA) attached to the spatial path, (c) feature maps obtained by CBAM (CPA) attached to the context path.

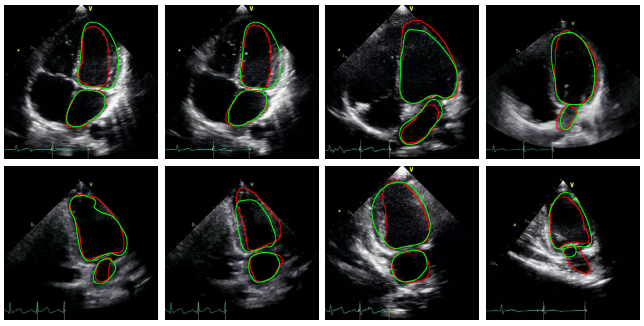


FIGURE 11. Images that do not perform well using the proposed method. The images at the first row are 4CH view, and those at the last row are 2CH view.

Although promising segmentation performance is achieved by the proposed AIDAN method, there are still limitations. The main limitation of our approach is that the relationship between adjacent echocardiography image frame in a cardiac cycle is not captured. Also, the training dataset is insufficient for semantic segmentation in the medical image analysis. Although data augmentation technology is adopted, we still cannot fully use deep neural networks for feature learning. Thus, we cannot build the network with deep architecture. Also, our method fails to segment some 4CH and 2CH view echocardiography images with low contrast and signal loss in Fig.11.

Of our research, we would like to use ConvLSTM [51] to learn the relationship between a cardiac cycle. Also, we would like to try different methods to combine multi-scale feature (e.g., U-Net++ [52] and CU-Net [53]), which seem a good choice. Also, we would like to try other backbone networks (e.g., DenseNet [22], FishNet [54]) for better feature extraction.

VI. CONCLUSION

In this paper, we present a novel end-to-end AIDAN with dual-path for pediatric echocardiography segmentation. Specifically, AIDAN contains two paths: the spatial path and the context path. The spatial path preserves low-level spatial information from pediatric echocardiography. The context path utilizes a backbone deep network and global average pooling to obtain a sizeable receptive field and

to capture high-level context semantic features efficiently. CBAM is used to guide the whole network to learn the most discriminative features. With the fusion of both low-level spatial details and high-level context semantic features, the proposed method outperforms related state-of-the-art segmentation methods in terms of Dice coefficient in pediatric echocardiography data.

REFERENCES

[1] R. A. Krasuski and T. M. Bashore, "Congenital heart disease epidemiology in the United States," *Circulation*, vol. 134, no. 2, pp. 110–113, Jul. 2016.

[2] S. Mendis, P. Puska, and B. Norrving, *Global Atlas on Cardiovascular Disease Prevention and Control*. Geneva, Switzerland: World Health Organization, 2011.

[3] X. Chen, S. Zhao, and X. Yang, "OC12.03: A national survey of fetal congenital heart diseases in China," *Ultrasound Obstetrics Gynecol.*, vol. 46, p. 26, Sep. 2015.

[4] L. Lopez, S. D. Colan, P. C. Frommelt, G. J. Ensing, K. Kendall, A. K. Younoszai, W. W. Lai, and T. Geva, "Recommendations for quantification methods during the performance of a pediatric echocardiogram: A report from the pediatric measurements writing group of the American society of echocardiography pediatric and congenital heart disease council," *J. Amer. Soc. Echocardiograph.*, vol. 23, no. 5, pp. 465–495, May 2010.

[5] J. D. Thomas, "Guidelines and recommendations for digital echocardiography: A report from the digital echocardiography committee of the American society of echocardiography," *J. Amer. Soc. Echocardiograph.*, vol. 18, no. 3, pp. 287–297, 2005.

[6] Q. Zheng, H. Delingette, N. Duchateau, and N. Ayache, "3-D consistent and robust segmentation of cardiac images by deep learning with spatial propagation," *IEEE Trans. Med. Imag.*, vol. 37, no. 9, pp. 2137–2148, Sep. 2018.

[7] O. Oktay, E. Ferrante, K. Kamnitsas, M. Heinrich, W. Bai, J. Caballero, S. A. Cook, A. De Marvao, T. Dawes, D. P. O'regan, B. Kainz, B. Glocker, and D. Rueckert, "Anatomically constrained neural networks (ACNNs): Application to cardiac image enhancement and segmentation," *IEEE Trans. Med. Imag.*, vol. 37, no. 2, pp. 384–395, Feb. 2018.

[8] W. Bai *et al.*, "Automated cardiovascular magnetic resonance image analysis with fully convolutional networks," *J. Cardiovascular Magn. Reson.*, vol. 20, no. 65, pp. 1–12, Sep. 2018.

[9] S. Leclerc, E. Smistad, J. Pedrosa, A. Ostvik, F. Cervenansky, F. Espinosa, T. Espeland, E. A. R. Berg, P.-M. Jodoin, T. Grenier, C. Lartizien, J. Dhooge, L. Lovstakken, and O. Bernard, "Deep learning for segmentation using an open large-scale dataset in 2D echocardiography," *IEEE Trans. Med. Imag.*, vol. 38, no. 9, pp. 2198–2210, Sep. 2019.

[10] G. Veni, M. Moradi, H. Bulu, G. Narayan, and T. Syeda-Mahmood, "Echocardiography segmentation based on a shape-guided deformable model driven by a fully convolutional network prior," in *Proc. IEEE 15th Int. Symp. Biomed. Imaging (ISBI)*, Apr. 2018, pp. 898–902.

[11] R. Shahzad, Q. Tao, O. Dzyubachyk, M. Staring, B. P. Lelieveldt, and R. J. Van Der Geest, "Fully-automatic left ventricular segmentation from long-axis cardiac cine MR scans," *Med. Image Anal.*, vol. 39, pp. 44–55, Jul. 2017.

[12] D. M. Vigneault, W. Xie, C. Y. Ho, D. A. Bluemke, and J. A. Noble, "Ω-net (omega-net): Fully automatic, multi-view cardiac MR detection, orientation, and segmentation with deep neural networks," *Med. Image Anal.*, vol. 48, pp. 95–106, Aug. 2018.

[13] M. A. Morales, D. Izquierdo-Garcia, I. Aganj, J. Kalpathy-Cramer, B. R. Rosen, and C. Catana, "Implementation and validation of a three-dimensional cardiac motion estimation network," *Radiol., Artif. Intell.*, vol. 1, no. 4, Jul. 2019, Art. no. e180080.

[14] G. Carneiro, J. C. Nascimento, and 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.*, vol. 21, no. 3, pp. 968–982, Mar. 2012.

[15] J. Zhang, "Fully automated echocardiogram interpretation in clinical practice: Feasibility and diagnostic accuracy," *Circulation*, vol. 138, no. 16, pp. 1623–1635, 2018.

[16] L. Yu, Y. Guo, Y. Wang, J. Yu, and P. Chen, "Segmentation of fetal left ventricle in echocardiographic sequences based on dynamic convolutional neural networks," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 8, pp. 1886–1895, Aug. 2017.

[17] L. Vargas-Quintero, B. Escalante-Ramírez, L. C. Marín, M. G. Huerta, F. A. Cosio, and H. B. Olivares, "Left ventricle segmentation in fetal echocardiography using a multi-texture active appearance model based on the steered Hermite transform," *Comput. Methods Programs Biomed.*, vol. 137, pp. 231–245, Dec. 2016.

[18] X. Yang, L. Yu, S. Li, X. Wang, N. Wang, J. Qin, D. Ni, and P.-A. Heng, "Towards automatic semantic segmentation in volumetric ultrasound," in *Medical Image Computing and Computer Assisted Intervention*. Cham, Switzerland: Springer, 2017, pp. 711–719.

[19] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[21] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[22] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4700–4708.

[23] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, Jun. 2019, pp. 3146–3154.

[24] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[25] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 7794–7803.

[26] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, Jul. 2017, pp. 2881–2890.

[27] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 7132–7141.

[28] J. Zhang, Y. Xie, Y. Xia, and C. Shen, "Attention residual learning for skin lesion classification," *IEEE Trans. Med. Imag.*, vol. 38, no. 9, pp. 2092–2103, Sep. 2019.

[29] C. Krittanawong, K. W. Johnson, R. S. Rosenson, Z. Wang, M. Aydar, U. Baber, J. K. Min, W. H. W. Tang, J. L. Halperin, and S. M. Narayan, "Deep learning for cardiovascular medicine: A practical primer," *Eur. Heart J.*, vol. 40, no. 25, pp. 2058–2073, Jul. 2019.

[30] D. Shen, G. Wu, and H. Suk, "Deep learning in medical image analysis," *Annu. Rev. Biomed. Eng.*, vol. 19, pp. 221–248, Jun. 2017.

[31] S. Liu, Y. Wang, X. Yang, B. Lei, L. Liu, S. X. Li, D. Ni, and T. Wang, "Deep learning in medical ultrasound analysis: A review," *Engineering*, vol. 5, no. 2, pp. 261–275, Apr. 2019.

[32] L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara, and D. Rueckert, "DRINet for medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 37, no. 11, pp. 2453–2462, Nov. 2018.

[33] Z. Lin, S. Li, D. Ni, Y. Liao, H. Wen, J. Du, S. Chen, T. Wang, and B. Lei, "Multi-task learning for quality assessment of fetal head ultrasound images," *Med. Image Anal.*, vol. 58, Dec. 2019, Art. no. 101548.

[34] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017.

[35] Q. Tao, W. Yan, Y. Wang, E. H. M. Paiman, D. P. Shamonin, P. Garg, S. Plein, L. Huang, L. Xia, M. Sramko, J. Tintera, A. De Roos, H. J. Lamb, and R. J. Van Der Geest, "Deep learning–based method for fully automatic quantification of left ventricle function from cine MR images: A multivendor, multicenter study," *Radiology*, vol. 290, no. 1, pp. 81–88, Jan. 2019.

[36] S. Gandhi, W. Mosleh, J. Shen, and C.-M. Chow, "Automation, machine learning, and artificial intelligence in echocardiography: A brave new world," *Echocardiography*, vol. 35, no. 9, pp. 1402–1418, Sep. 2018.

[37] X. Du, R. Tang, S. Yin, Y. Zhang, and S. Li, "Direct segmentation-based full quantification for left ventricle via deep multi-task regression learning network," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 3, pp. 942–948, May 2019.

[38] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr, "Conditional random fields as recurrent neural networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1529–1537.
[39] V. Koltun, "Efficient inference in fully connected CRFS with Gaussian edge potentials," in *Proc. Adv. Neural Inf. Process. Syst.*, Granada, Spain, 2011, pp. 109–117.
[40] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention*. Cham, Switzerland: Springer, 2015, pp. 234–241.
[41] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "BiSeNet: Bilateral segmentation network for real-time semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, Munich, Germany, 2018, pp. 1–17.
[42] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "CBAM: Convolutional block attention module," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, Munich, Germany, 2018, pp. 3–19.
[43] J. Duan, "Automatic 3D bi-ventricular segmentation of cardiac images by a shape-refined multi-task deep learning approach," *IEEE Trans. Med. Imag.*, vol. 38, no. 9, pp. 2151–2164, Jan. 2019, doi: 10.1109/TMI.2019.2894322.
[44] R. Girdhar, D. F. Fouhey, M. Rodriguez, and A. Gupta, "Learning a predictable and generative vector representation for objects," in *Proc. Eur. Conf. Comput. Vis.* Amsterdam, The Netherlands: Springer, 2016, pp. 484–499.
[45] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.* Amsterdam, The Netherlands: Springer, 2016, pp. 483–499.
[46] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "BiSeNet: Bilateral segmentation network for real-time semantic segmentation," in *Computer Vision—ECCV*, vol. 11217. Cham, Switzerland: Springer, 2018, pp. 334–349.
[47] Y. Hu, L. Guo, B. Lei, M. Mao, Z. Jin, A. Elazab, B. Xia, and T. Wang, "Fully automatic pediatric echocardiography segmentation using deep convolutional networks based on BiSeNet," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Berlin, Germany, Jul. 2019, pp. 6561–6564.
[48] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters–improve semantic segmentation by global convolutional network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4353–4361.
[49] B. Steiner, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2019, pp. 8024–8035.
[50] A. Sergeev and M. Del Balso, "Horovod: Fast and easy distributed deep learning in TensorFlow," 2018, *arXiv:1802.05799*. [Online]. Available: https://arxiv.org/abs/1802.05799
[51] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. Adv. Neural Inf. Process. Syst.*, Montréal, QC, Canada, 2015, pp. 802–810.
[52] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Granada, Spain: Springer, 2018, pp. 3–11.
[53] Z. Tang, X. Peng, S. Geng, L. Wu, S. Zhang, and D. Metaxas, "Quantized densely connected U-Nets for efficient landmark localization," in *Proc. Conf. Eur. Comput. Vis.*, 2018, pp. 339–354.
[54] S. Sun, J. Pang, J. Shi, S. Yi, and W. Ouyang, "FishNet: A versatile backbone for image, region, and pixel level prediction," in *Proc. Adv. Neural Inf. Process. Syst.*, Montréal, QC, Canada, 2018, pp. 754–764.

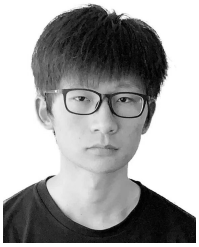
YUJIN HU received the bachelor's degree in biomedical engineering from South-Center University for Nationalities, in 2013. He is currently the master's degree with Shenzhen University. His research interest includes segmentation and analysis in biomedical images.

BEI XIA received the bachelor's degree from the Jianghan University Medical College, in 1984. She is currently the Director of the Department of Ultrasonography, Shenzhen Children's Hospital, China. She is also a Master Supervisor of China Medical University, and the Shantou University Medical College. Her research interests include congenital heart disease, Kawasaki disease, hepatoblastoma, and developmental Dysplasia of the Hip.

MUYI MAO graduated from Shantou University Medical College, in 2018. He is currently a Resident with the Ultrasound Department, Shenzhen Children's Hospital, China. His current research interests include ultrasonography and radiology.

ZELONG JIN received the bachelor's degree from China Medical University, in 2018. He is currently pursuing the master's degree with the Department of Ultrasonography, Shenzhen Children's Hospital, China. His research interests include heart disease and hematological disease.

JIE DU received the B.S. degree in computer science and technology from the University of Shihezi, Xinjiang, China, in 2013, and the Ph.D. degree from the Department of Computer and Information Science, University of Macau, Macau, China, in 2019. She is currently an Assistant Professor with the Health Science Center, School of Biomedical Engineering, Shenzhen University, Shenzhen, China. Her research interests include machine learning methods and medical image processing.



LIBAO GUO graduated from Nanchang Hangkong University. He is currently pursuing the master's degree with the School of Biomedical Engineering, Shenzhen University. His research interest is medical image processing.

ALEJANDRO F. FRANGI (Fellow, IEEE) received the B.Sc. and M.Sc. degrees in telecommunications engineering from the Technical University of Catalonia, Spain, in 1996, and the Ph.D. degree in biomedical imaging from the Image Sciences Institute, Utrecht University, in 2001. He is the Diamond Jubilee Chair of Computational Medicine at the University of Leeds, Leeds, U.K., with joint appointments at the School of Computing and the School of Medicine. His main research interests are in medical image computing, medical imaging, and image-based computational physiology.

BAIYING LEI (Senior Member, IEEE) received the M.Eng. degree in electronics science and technology from Zhejiang University, China, in 2007, and the Ph.D. degree from Nanyang Technological University (NTU), Singapore, in 2013. She is currently an Associate Professor with the School of Biomedical Engineering, Shenzhen University, China. Her current research interests include medical image analysis, machine learning, digital watermarking, and signal processing.

TIANFU WANG received the Ph.D. degree in biomedical engineering from Sichuan University, in 1997. He is currently a Professor with the School of Biomedical Engineering, and the Associate Chair of the Health Science Center, Shenzhen University, China. His research interests include ultrasound image analysis, medical image processing, pattern recognition, and medical imaging.