MCAL: An Anatomical Knowledge Learning Model for Myocardial Segmentation in 2-D Echocardiography

Xiaoxiao Cui, Pengfei Zhang, Yujun Li, Zhi Liu, Xiaoyan Xiao, Yang Zhang, Longkun Sun, Lizhen Cui, *Member, IEEE*, Guang Yang, *Senior Member, IEEE*, and Shuo Li, *Senior Member, IEEE*

***Abstract*—Segmentation of the left ventricular (LV) myocardium in 2-D echocardiography is essential for clinical decision making, especially in geometry measurement and index computation. However, segmenting the myocardium is a time-consuming process and challenging due to the fuzzy boundary caused by the low image quality. The ground-truth label is employed as pixel-level class associations or shape regulation in segmentation, which works limit for effective feature enhancement for 2-D echocardiography. We propose a training strategy named multiconstrained aggregate learning (referred to as MCAL), which leverages anatomical knowledge learned through ground-truth labels to infer segmented parts and discriminate boundary pixels. The new framework encourages the model to focus on the features in accordance with the learned anatomical representations, and the training objectives incorporate a boundary distance transform weight (BDTW) to enforce a higher weight value on the boundary region, which helps to improve the segmentation accuracy. The proposed method is built as an end-to-end framework with a top-down, bottom-up architecture with skip convolution fusion blocks and carried out on two datasets (our dataset and the public CAMUS dataset). The comparison study shows that the proposed network outperforms the other segmentation baseline models, indicating that our method is beneficial for boundary pixels discrimination in segmentation.**

***Index Terms*—Boundary distance transform weight (BDTW), multiconstrained aggregate learning (MCAL), myocardial segmentation.**

Manuscript received December 1, 2021; accepted February 11, 2022. Date of publication February 15, 2022; date of current version March 28, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 91846205, in part by the Joint Fund for Smart Computing of the Shandong Natural Science Foundation under Grant ZR2020LZH013, in part by the Major Fundamental Research of the Natural Science Foundation of Shandong Province under Grant ZR2019ZD05, in part by the Fundamental Research Funds of Shandong University under Grant 2018JC009, in part by the General Financial Grant from the China Postdoctoral Science Foundation under Grant 2017M622215, in part by the MRC under Grant MC/PC/21013, and in part by the UKRI Future Leaders Fellowship under Grant MR/V023799/1. *(Xiaoxiao Cui and Pengfei Zhang contributed equally to this work.) (Corresponding authors: Yujun Li; Zhi Liu; Lizhen Cui.)*

Xiaoxiao Cui, Yujun Li, and Zhi Liu are with the School of Information Science and Engineering, Shandong University, Qingdao 266237, China (e-mail: liyujun@sdu.edu.cn; liuzhi@sdu.edu.cn).

Pengfei Zhang and Longkun Sun are with the Department of Cardiology, Qilu Hospital, Cheeloo College of Medicine, Shandong University, Jinan, Shandong 250012, China.

Xiaoyan Xiao is with the Department of Nephrology, Qilu Hospital, Shandong University, Jinan 250012, China.

Yang Zhang is with the Department of Radiology, Qilu Hospital, Shandong University, Jinan, Shandong 250012, China.

Lizhen Cui is with the Joint SDU-NTU Center for Artificial Intelligence Research (C-FAIR), Shandong University, Jinan 250101, China (e-mail: clz@sdu.edu.cn).

Guang Yang is with the National Heart and Lung Institute, Imperial College London, London SW7 2AZ, U.K.

Shuo Li is with the Department of Medical Imaging, Western University, London, ON N6A 3K7, Canada.

Digital Object Identifier 10.1109/TUFFC.2022.3151647

I. Introduction

ECHOCARDIOGRAPHY is routinely used in the diagnosis and management of cardiovascular disease because it can provide real-time images of a beating heart, combined with its availability and portability [1]. Heart function assessment, such as diastolic analysis, the calculation of the cardiac output, and the ejection fraction (EF), are key determinants of clinical decisions. The segmentation of the left ventricular (LV) myocardium helps accurate quantification of these indexes in the clinical workflow. Thus, developing an automatic approach for accurate myocardial segmentation liberates radiologists from manual annotation.

Several research works have been performed efficiently on the segmentation in B-mode echocardiography in the past few decades [2]–[4]. With the combination of various feature enhancement modules [5], [6] and different deep network architectures [7]–[12], the ground truth is applied as a class associate or shape regulation by minimizing the loss function. However, these methods still have scope for improvement. First, methods that focus on feature enhancement to achieve a better result still work limit for echocardiography since the limitations of the ultrasound image due to resolution, the presence of speckle noise, and artifacts caused by the complex interaction between the tissue and ultrasound usually lead to an ambiguous border between the myocardium and chamber [see Fig. 1(a)], making it difficult for an accurate delineation of the myocardium [see Fig. 1(b)] by feature enhancement modules. Second, works that regulate the segmented output with some constraint strategy [9], [10] are much like postprocession and global constraint. However, boundary pixels of echocardiography are hard to capture by shape constraints because imaging quality varies from subject to subject [see

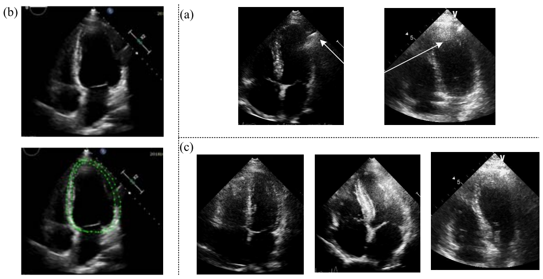


Fig. 1. Typical images extracted from our dataset. (a) Examples of samples with an ambiguous border (left: fuzzy chamber border; right: fuzzy apical border). (b) Illustration of the annotation of the myocardium. Top: original ultrasound image. Bottom: annotation of the myocardium with a green dot. (c) Different image quality (left: good image quality; middle: medium image quality; and right: poor image quality).

Fig. 1(c)], giving rise to difficulty in capturing the intensity change on the boundary.

To fully use the annotations to address the limitations, we propose a novel training strategy named multiconstrained aggregate learning (MCAL). Specifically, we force the distributions divergence of the latent features of the input and the ground truth to be close in the training process. This helps to infer anatomical structure from the deeper layer of the encoder. Further adjusting the scale and offset of the learned anatomical information by a featurewise linear modulation (FiLM) [13], the segmented relevance feature information is enhanced. Finally, upon observing that the boundary is hard to detect, we further enhance a higher weight on the border neighborhood pixels by proposing a boundary distance transform weight (BDTW) for the segmentation loss, which acts as guidance for penalizing the learning process.

The main contributions of our proposed framework are given as follows.

1) Our method derives segmented-relevance information by narrowing distribution divergence between the latent space of input and label, which exploits anatomical knowledge to guide feature enhancement. FiLM is applied to enhance segmented-relevant features with the guidance of the anatomical information, which restrains the irrelative features under low image quality.
2) A novel BDTW is applied to the cross-entropy loss. It forces the network to focus on the boundary region pixels in each training batch and improves the discrimination on boundary pixels, which is useful in cases with low image quality.

II. Related Works

There have been many works on the segmentation of B-mode echocardiography, which mainly falls into two categories: the traditional methods and the deep learning methods. Most of the solutions based on traditional methods need prior information, such as the appearance or shape of the LV [2], [4], [14]–[16], which presents an assumption that the border between myocardium and blood pool is accessible and, therefore, possible to achieve good segmentation results based on prior knowledge. As these studies are dependent on predefined knowledge, so they may fail if data vary from the information stored in the priors. The other methods aim to minimize the energy function by tuning a large number of parameters [3], [17], [18].

Recently, with the development of deep learning in medical image analysis [19], [20], segmentation methods based on the deep convolutional neural network (CNN) learned the features with different convolutional kernels and connection methods to obtain accurate and robust results [21]–[26]. Two publicly available datasets in echocardiography CAMUS [27] and Dynamic-Echonet [28] are researched, which proved that the deep learning algorithm outperformed in the tasks of segmenting the left ventricle, especially the encoder–decoder-based architectures [27]. Several works deal with echocardiographic sequence segmentation by incorporating temporal information, such as optical flow [29] and hierarchical convolution aggregated with temporal relevance [30]. However, the temporal information may deteriorate significantly in a low-quality frame because of the high noise. With insights from shape regularization on the prediction, the anatomically constrained neural networks (ACNNs) [9] and shape reconstruction neural network [31] have worked to maintain a realistic shape of the resulting segmentation without postprocession.

VAE [32] approximates posterior distribution via a parameterized variational inference. The distribution is enforced to be close to a normal distribution as a regularization, which is applied in the cross-modality image segmentation [33], [34]. By regularization, the model can learn a shared domain-invariant latent space with the same distribution. In this work, we applied regularization to narrow the distribution divergence between the latent features of the input and the label, which helps to infer the segmented information.

III. Methodology

In this article, the MCAL leverages anatomical and spatial knowledge learned through ground-truth labels on the myocardial segmentation in 2-D echocardiography on the backbone of an encoder–decoder architecture, as shown in Fig. 2. A latent representation encoder maps the input to a latent space by learning the high-level semantical information. For a raw input image, the spatial space contains the segmented anatomical information mixed with the other contexture. Thus, we apply the Kullback–Leibler (KL) divergence loss to learn the distribution difference between the spatial space and the anatomical space of the ground-truth label. On the other hand, FiLM highlights the feature responses in relevant segmentation regions by modulating the spatial space. Furthermore, the skip convolution fusion blocks are applied to discriminate more fine-grained details from the intermediate feature maps through the encoder. Finally, the BDTW focuses on the border neighborhood pixels in each training batch and improves the segmentation accuracy. Fig. 3 shows the detailed architecture of each module.

A. Anatomical Information Derivation With Distribution Divergence Regularization

VAE is a generated model based on samples from a latent variable, of which the posterior distribution is approximated from the input. For an input x, the approximate posterior distribution of its latent variable z can be estimated by an encoder $p(\cdot|\cdot)$. The encoded distributions are set to be an isotropic

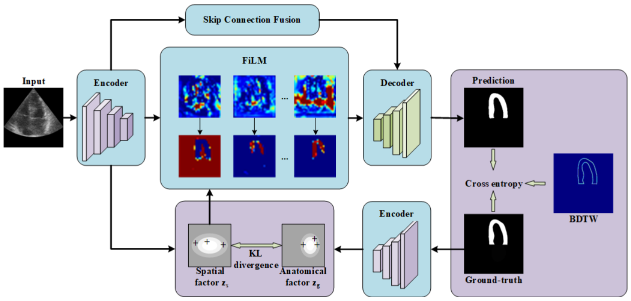


Fig. 2. Block diagram of the proposed model. An input image and its corresponding annotation are encoded to a spatial space *s* and anatomical space *g*, respectively, using an encoder f_s and f_g. Then, *s* and the spatial factor z_s are combined as an input to a decoder f_h to produce a myocardial segmentation prediction. The spatial factor z_s is constrained to learn the distribution, which is close to that of anatomical factor z_g. The parameters of the whole model (the black line and the red line) propagate to achieve an optimal result in the training stage, and the parameters among the black lines are loaded in the test stage.

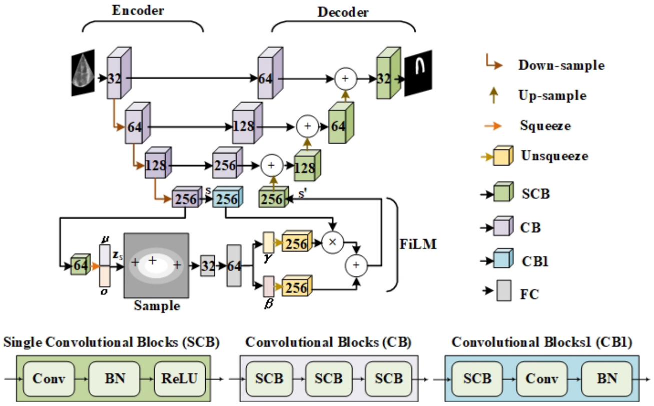


Fig. 3. Architectures of the encoder–decoder that makes up the MCAL network. The spatial encoder module is constructed of four convolutional blocks and produces spatial space *s* for the input image. Then, it is modulated by the spatial factor z_s with an FiLM. Finally, the decoder combines the bridge layer from the encoder with a skip convolution fusion block to produce a segmentation prediction of the myocardial. The number on each box represents the channels of feature maps. Here, the "Conv," "BN," "ReLU," and "FC" represent the convolution layer, the batch normalization layer, the rectified linear unit activation layer, and the fully connected layer, respectively.

multivariate Gaussian $N(\mu, \sigma)$ with mean μ and variance σ. Specifically, the encoded feature space produces dimensional mean and diagonal covariance by dimensional squeeze, and then, they are sampled to be an axis-aligned Gaussian distribution to yield the final latent variable. A decoder $q(\cdot|\cdot)$ converts the samples from z back to the input space.

Motivated by the distribution regularization in VAE, we aim to apply the regularization to narrow distribution divergence between the latent feature space of input and label. Specifically, we first transform the input and label into a latent feature variable by a latent representation encoder, separately. Then, we estimated the approximate posterior distribution by a parameterized variational form. The divergence of the distribution from the input and label is regularized to infer segmented-relevance information in segmentation.

The latent representation encoder transforms the input into a spatial representation in our model. The encoded spatial representation is a group of feature maps that contain the spatial information of the input in different channels, so we define spatial feature space s as $f_s(x)$, specifically for an input x. Considering that the encoded latent space of the annotation ground truth y mainly includes the anatomic information, we define the anatomical feature space g as $f_g(y)$. The latent factor from the encoded feature space is denoted as z_s and z_g for the input and ground truth, respectively. Their corresponding approximate probability distribution is denoted as $p_\theta(z_s|x, s)$ and $p_\phi(z_g|y, g)$, respectively. The distance between $p_\theta(z_s|x, s)$ and $p_\phi(z_g|y, g)$ is then used as an effective regularization for segmentation directly. The distribution discrepancy convergences gradually during the network training. The distribution difference is directly penalized by the KL divergence

$$L_{\text{kl}} = D_{\text{KL}}\big(p_\theta(z_s|x, s)||p_\phi(z_g|y, g)\big). \tag{1}$$

B. Feature Enhancement With FiLM

An FiLM constrains the information stored in the spatial space by adjusting the scale and offset of the sampled data over the spatial factor z_s. The rescale and offset coefficients are predicted from spatial factor s, which, after a series of convolutions, are conditioned by z_s samples. Specifically, z_s is sampled and then fed into two fully connected layers to obtain the scale γ and offset β, as shown in Fig. 3. To modulate each feature map in the spatial space s, γ, and β are unsqueezed by the dimensional expansion. Then, the spatial space s is passed through a convolution layer, and the modulated output of each channel is formulated by an elementwise multiplication $\odot$ and addition operation as follows:

$$F_c' = \gamma_c \odot F_c + \beta_c. \tag{2}$$

Each feature map is affined to learn from the sampled data; here, F_c represents the feature map in each channel c.

To verify the effectiveness of FiLM, the feature maps of some selected channels relevant to the segmentation before and after FiLM are visualized in Fig. 4. The first column represents the input images and predicted segmentation results of the MCAL. The feature maps in the first and third rows demonstrate that the spatial space contains anatomical information drowned in the complicated semantical and spatial information. The second and fourth rows display the corresponding output of the FiLM. It is obvious that, with the modulation of the spatial factor, the network has learned critical information of the segmented structure in some channels. However, we observed that the feature map of channel 12th is weakly associated with the myocardial segmentation, and the structure in the 27th and 87th channels is incomplete. Thus, we use the skip convolution fusion block, which combines the semantic information of the encoder and the relative rich anatomical information in the decoder to solve the problem and fulfill segmentation.

C. Segmentation With Decoder

The architecture of the decoder shown in Fig. 3 is a bottom-up structure with a skip connection with the encoder.

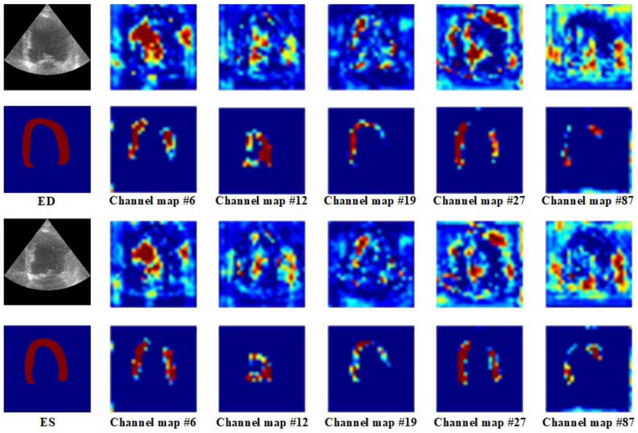


Fig. 4. Visualization of the learned feature maps of some selected channels before and after FiLM. The input images and the corresponding prediction results of our method are in the first column. The five most relevant channel maps of the spatial information before and after modulation are in the first and third rows, and the second and fourth rows, respectively.

Each upsampling layer adds the feature maps of the bridge pathway that makes input of the single convolutional block (SCB). The bridge pathway between the encoder and the decoder consists of a convolutional block with three successive convolution layers.

Formally, we formulate the bridge pathway as follows: let F_i^e and F_i^d denote the learned feature maps with the same size before the ith downsampling layer along with the encoder and the ith upsampling layer along with the decoder, respectively. The stack of feature maps represented by F_i^s is computed as

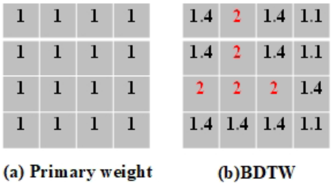
$$F_i^s = H_1\left(F_i^e\right) + U\left(F_{i+1}^d\right) \tag{3}$$

where function $H_1(\cdot)$ is a convolutional block operation and $U(\cdot)$ denotes an upsampling layer. Then, we obtain the output by $F_i^d = H_2 F_i^s$, where function $H_2(\cdot)$ is a convolution operation followed by batch normalization and ReLU activation.

The decoder recovers the features in each upsampling layer by fusing corresponding semantical feature information of the encoder. Specifically, the modulated spatial output s' is first upsampled by a bilinear operation, which restores the dimension of the feature maps gradually in each upsample layer and, finally, achieves pixel-level prediction. Mainly, the skip convolution introduces the feature maps of the encoder selectively, which contains more semantic information closer to that of the feature maps in the decoder. Then, outputs of the skip convolution fusion block are fused with the previous bilinear layer output of the lower skip convolution fusion block. The fusion of the enhanced semantical feature maps with the same size between the encoder and the decoder by the skip convolution fusion block can facilitate optimizer during network training.

D. Boundary Distance Transform Weight Loss

Some works have proposed to solve the label imbalance by multiplying the class weight and tuning the weight value of hard examples iteratively [35]. The boundary is hard to detect in segmentation, especially for ultrasound images with artifacts and speckle noise. Errors located on boundaries affect further index calculation and analysis. A simple and straightforward way to solve this problem is to assign higher weights to the adjacent pixels of the boundary. More weight assigned only to the boundary pixels leads the network to strengthen the boundary information. However, determining the weight value of the boundary region is difficult.



1	1	1	1
1	1	1	1
1	1	1	1
1	1	1	1

(a) Primary weight

1.4	2	1.4	1.1
1.4	2	1.4	1.1
2	2	2	1.4
1.4	1.4	1.4	1.1

(b)BDTW

Fig. 5. Illustrations of (a) primary weight and (b) BDTW; the weight values are different according to the distance to the boundary pixels (the red). The BDTW assigns a higher weight to the boundary region, which directly learns the desirable result and, therefore, helps reduce the prediction errors.

In this article, we apply the boundary information to calculate the new weight, which displays the distance of each pixel to the boundary. The weighted distance transform map is decreased exponentially according to the Hausdorff distance (HD) to the boundary. It forces the network to pay more attention to the boundary, which is formulated as follows:

$$W_{i,j} = \exp(-\mathrm{HD}[i,j]) = \exp\left(-\min_{(k,l)\in q} d([i,j],[k,l])\right) \tag{4}$$

where d is the distance of each pixel $[i,j]$ of the ground truth to $[k,l]$, which belongs to a boundary set q. The standard Euclidean distance $d([i,j],[k,l]) = \sqrt{(k-i)^2+(l-j)^2}$ is used to calculate the distance between pixels. As shown in Fig. 5(b), the BDTW assigns the pixels of the boundary region to higher values. Such a mechanism penalizes the hard prediction boundary pixels and, therefore, helps to reduce the overall prediction errors.

Finally, the BDTW loss is obtained by

$$L_{bdtw} = \left(\lambda W_{i,j} + 1\right) \odot L_{\mathrm{ce}} \tag{5}$$

where λ is a hyperparameter, L_{ce} is the cross-entropy, and $\odot$ is the Hadamard product. Since the weight of pixels that are far from the boundary is small, hence, to mitigate the vanishing gradient issue, all the weight value is increased by 1. Moreover, the BDTW can be computed in the dataset only once, which does not burden the calculations.

We also adopt the Lovasz-softmax loss [36] as the loss function to measure the result of the segmentation, as the framework fulfills the segmentation based on the pixel classification problem. The Lovasz-softmax loss function L_{ls} directly optimizes the mean intersection-over-union loss in the context of semantic image segmentation

$$L_{ls} = \frac{1}{|C|}\sum_{c\subset C}\overline{\Delta_{J_c}}(m(c)) \tag{6}$$

where Δ_{J_c} is the loss surrogate, $m(c)$ corresponds to the vector of pixel identification errors, and C is the class number.

In general, the final loss function is the weighted sum of the segmentation loss and KL loss, and the trainable parameters

θ_w are regulated with the L2 paradigm by factor η, which is formulated as

$$L = \lambda_l L_{\mathrm{ls}} + L_{bdtw} + \lambda_v L_{\mathrm{kl}} + \eta \|\theta_w\|_2^2 \tag{7}$$

where λ_l and λ_v are hyperparameters to allocate the corresponding loss.

IV. Material and Experimental Results

A. Dataset Information and Annotations

1) Our Echocardiography Dataset: We evaluated the proposed MCAL on our dataset, which contains a total of 1472 frames of 11 healthy subjects, and is collected from two hospitals with different devices by Philips and GE, with ethics approval from the Clinical Medical Research Ethics Board. The privacy information of patients is erased at the workstation. The temporal rate is 65–70-Hz among frames. The pixel resolution of images from the devices is 0.353×0.353 mm^2. We research the apical four-chamber view (A4C) of each examination of those subjects in the experiment. From these images, we randomly select these subjects to train the model and test. Considering the intersubjective appearance difference and the intrasubjective frame relevance, we split the dataset into a cross-validation set for training and a dependent test set with a ratio of 9:2 on a subject basis, based on a ratio of 8:2 on an image basis, to illustrate the robustness of our proposed method. Besides, we abandon the first and last frames of the echo cine due to poor image quality. Each subject contains at least one temporally cropped sequence that captures one complete cardiac cycle from ES to ED. An expert annotates the myocardium of each image in the dataset manually according to [1]. The other expert confirms the inconsistency of the annotation and the labels. The annotation masks are considered as ground truth to train our model.

2) CAMUS Dataset: To comprehensively evaluate the performance of our model, we also use the public CAMUS dataset for verification in the experiment. The CAMUS dataset contains 500 patients with an apical two-chamber view (A2C) and an A4C view, acquired from a single vendor and center. The pixel resolution of the image is 0.154×0.154 mm^2. Only the annotations of the ES and ED frames are available. Because the annotations of the final 50 are not given in the training data, we adopted tenfold cross-validation for the evaluation on the CAMUS dataset on the left 450 patients.

B. Data Prepossessing

The raw image is preprocessed to keep the cardiac part only before feeding into the model. We randomly apply rotation augmentation to avoid overfitting, and the rotation angle is between $-5°$ and $5°$ randomly according to the real echo cine. These images are resized to 224 × 224, and the gray value has been normalized to the range [0, 1]. For the comparison study and ablation study on the CAMUS with different views and phases, we apply the same prepossession as [37]. Please refer to [37] for a more detailed preprocession of the CAMUS dataset. Since we preprocessed the images by resizing the image to 224 × 224, the pixel distance in the resized image is scaled down from the original image. We calculated the distance metrics by multiplying the rescaled pixel distance specified on the resized image. More importantly, we set the length and width of the image to be the same by filling zeros before resizing the image. Thus, the aspect ratio of pixel distance is unchanged before and after the preprocessing.

C. Experimental Setup

We use tenfold cross-validation to train the model. Since our model is based on an encoder–decoder backbone with skip concatenate fusion, we adopted U-net with the same architecture design in [26]. In detail, the number of filters is the same as the U-net in the encoder and the decoder. We set the number of latent vectors to 32 due to the computation efficiency. The normal distribution is applied to initialize the parameters of the network at first. The Adam optimizer applies an initial learning rate of 0.0001 and a weight decay of 0.9 in initialization. The batch size was set to 7 for our dataset. For each fold cross-validation, we trained 100 epochs. The Dice coefficient [38] is used to assess the accuracy of the segment model. We performed our model on an NVIDIA GeForce RTX 2080Ti GPU in Pytorch.

We explored the impact of hyperparameters of the loss function on the behavior of MCAL. Because L_{ls} and L_{ce} are fundamental in the segmentation loss, we set λ_l to be 1 in our experiment. In addition, λ and λ_v are set the same value to evaluate their performance on segmentation. The performance of MCAL was evaluated under different parameter settings $\lambda \in \{1, 3, 5, 10, 15\}$. The results are shown in Table III. Totally, the best performance is obtained when $\lambda = 15$. We also observed that the Dice dropped obviously when $\lambda \in \{3, 5\}$, which was the worst performance. Table I illustrates that, although the performance varies when $\lambda \in \{1, 10, 15\}$, the magnitude of the variation is not very large, so the value of $\lambda = 10$ is adopted for its second-best performance among the three. We trained and test the model with the same parameters settings on our dataset.

D. Comparison With Existing Methods

1) Comparison on Our Dataset: The comparison study is carried out to evaluate the effectiveness of the network. We compared our method with the UNet, ACNN, and the effectiveness of BDTW on our dataset in this article. We used the geometrical metric for a comprehensive evaluation of the method: three area error metrics (precision, recall, and Dice) and two distance error metrics (absolute surface distance (ASD) and HD). The mean and standard deviation values of each metric were obtained from the cross-validating on the test dataset. We selected the best model on each fold validation set for the test.

Table II shows the experimental results on our echo dataset. The mean and the standard deviation values are used for each metric to perform the cross-validation procedure. The bold font indicates the best results for each metric. We observed that our framework outperforms other methods on all metrics, achieving the highest mean values of precision (75.43%) and Dice (76.16%), the lowest mean values of HD (5.85 mm), and significantly lower standard deviations of all metrics,

TABLE I

SEGMENTATION PERFORMANCE UNDER DIFFERENT HYPERPARAMETER SETTINGS ON CAMUS. BOLD NUMBERS REPRESENT THE BEST RESULTS OBTAINED

Weight	A2C			A4C		
	Dice(%)	d_m(mm)	d_H(mm)	Dice(%)	d_m(mm)	d_H(mm)
1	84.15±6.70	0.97±0.44	4.34±3.34	84.69±6.45	0.77±0.31	3.63±3.14
3	83.74±7.16	0.98±0.43	4.70±3.83	83.63±7.05	0.81±0.35	4.23±3.84
5	83.96±6.67	1.00±0.61	4.75±4.03	83.83±6.95	0.80±0.35	4.17±3.85
10	84.27±6.40	0.97±0.48	**4.33±3.38**	84.78±6.58	0.76±0.33	**3.40±2.57**
15	**84.39±6.60**	**0.95±0.42**	4.36±3.54	**84.92±6.50**	**0.75±0.30**	3.46±2.89
	ED			ES		
1	83.93±6.44	0.86±0.42	3.97±3.06	84.84±7.18	0.86±0.36	3.97±3.43
3	83.26±6.67	0.88±0.40	4.36±3.53	84.10±7.49	0.91±0.40	4.55±4.10
5	83.38±6.72	0.89±0.54	4.46±3.82	84.42±6.87	0.90±0.47	4.45±4.08
10	84.13±6.32	0.84±0.44	**3.88±2.95**	84.93±6.64	0.87±0.40	**3.81±3.09**
15	**84.22±6.41**	**0.83±0.38**	3.95±3.12	**85.11±6.67**	**0.85±0.37**	3.83±3.37

TABLE II

MCAL OUTPERFORMS THE OTHER METHODS UNDER DIFFERENT CONFIGURATIONS ON OUR DATASET. BOLD NUMBERS REPRESENT THE BEST RESULTS OBTAINED

Method	Precision (%)	Recall (%)	Dice (%)	d_H (mm)	d_m (mm)
UNet	69.20 ±7.11	73.62 ±6.13	71.12 ±5.31	7.21 ±4.32	0.80 ±0.16
ACNN	68.58 ±6.54	76.05 ±8.15	71.76 ±5.33	13.27 ±6.42	0.90 ±0.45
MCAL (w/o BDTW)	69.62 ±6.52	**80.44** **±4.54**	74.42 ±4.19	7.19 ±3.14	**0.70** **±0.09**
MCAL	**75.43** **±6.24**	76.18 ±7.00	**76.16** **±4.16**	**5.85** **±1.74**	0.83 ±0.21

especially with the BDTW. This finding demonstrates that a combination of shape and latent anatomical information brings improvement in myocardial segmentation. Fig. 6 presents some typical segmentation results, which visually illustrates that the mentioned method obtains a more accurate segmentation; especially, the BDTW keeps more fine anatomical information on the prediction.

All echocardiographic sequences from ES to ED are analyzed in Fig. 7 to observe the temporal performance of the proposed method. The precision, Dice, HD, and ASD are computed at each frame of a whole cardiac cycle of one test subject to assess the temporal stability. As shown in Fig. 7, all the four metrics fluctuate moderately between each frame in the entire cardiac cycle, which means that the proposed method has a limitation on a single image without spatial information. This limitation could be improved by taking into account the relevance of the successive frame.

2) Comparison on CAMUS: Since the domain gap in our dataset may affect the performance, we conducted another comparison experiment on the public CAMUS dataset to evaluate our method intuitively. For the CAMUS dataset, we compared the method with UNet++ [23], SegNet [40], CPFNet [39], HarDNet-MSEG [41], and PLANet in [37], except UNet and ACNN. While the first two are leading methods, the middle two are newly proposed public methods, and the last is a new method proposed on CAMUS. The Dice, ASD, and HD are used in the comparison study.

Geometrical results are analyzed comprehensively by performing the comparison study on the public CAMUS from the view and phase perspective to assess the influence of the latent representations in the myocardial segmentation between different training views. We carried out the comparison study without any postprocession, such as filling the hole and

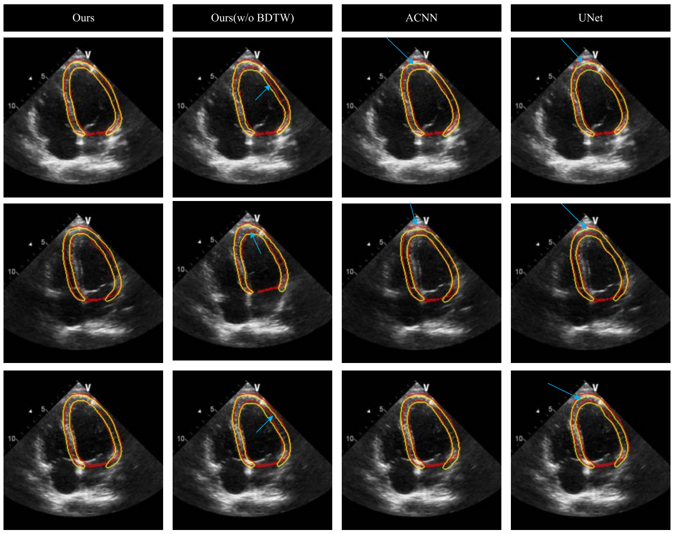


Fig. 6. Qualitative comparison of the results under four different settings on myocardial segmentation. Red and yellow denote the ground truth and predict, respectively. The blue arrow indicates the wrong prediction of the boundary region. The MCAL could improve the prediction accuracy, and the precision improves more for the pixels of the boundary region, revealing that the BDTW is especially effective for the boundary region.

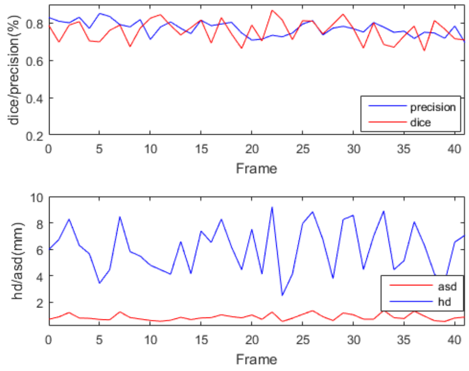


Fig. 7. Precision, Dice, HD, and ASD at different frames of the cardiac cycle of one test subject.

TABLE III
PERFORMANCE COMPARISON OF MCAL AGAINST EXISTING METHODS ON THE CAMUS DATASET. BOLD NUMBERS REPRESENT THE BEST RESULTS OBTAINED

Methods	A2C			A4C		
	Dice(%)	d_m(mm)	d_H(mm)	Dice(%)	d_m(mm)	d_H(mm)
Ours	85.33±5.65	0.92±0.37	**3.54±2.45**	**85.85±5.59**	0.73±0.30	3.61±2.98
CPFNet [39]	**85.84±6.70**	**0.84±0.41**	4.34±3.28	85.25±6.65	0.75±0.30	3.17±2.08
SegNet [40]	83.37±7.39	0.95±0.48	5.82±4.77	83.45±7.58	0.79±0.39	6.19±5.92
PLANet [37]	83.54±6.38	1.06±0.51	4.36±3.14	85.85±5.68	**0.71±0.32**	**2.97±2.07**
HarDNet-MSEG [41]	82.41±6.86	1.14±0.69	4.83±3.41	82.57±7.13	0.89±0.47	4.05±2.99
ACNN [9]	84.31±6.60	0.96±0.57	4.46±3.60	84.23±6.60	0.78±0.37	3.79±3.29
UNet [26]	79.84±8.53	1.28±0.95	6.74±5.10	81.50±7.74	0.91±0.48	5.97±5.05
UNet++ [23]	80.22±8.36	1.27±0.99	7.10±5.46	81.19±7.71	0.94±0.48	6.45±5.53
	ED			ES		
Ours	**85.10±5.58**	0.83±0.36	**3.42±2.28**	**86.08±5.59**	**0.81±0.38**	4.01±3.43
CPFNet [39]	85.08±6.56	**0.78±0.35**	3.94±2.86	86.00±6.77	0.85±0.34	3.27±2.25
SegNet [40]	82.97±7.13	0.86±0.46	6.13±5.43	83.86±7.80	0.88±0.43	5.88±5.31
PLANet [37]	83.68±6.16	0.92±0.51	4.14±3.25	85.71±5.96	0.85±0.40	**3.18±2.02**
HarDNet-MSEG [41]	81.81±6.87	1.03±0.71	4.72±3.54	83.16±7.06	1.00±0.47	4.16±2.85
ACNN [9]	83.81±6.56	0.87±0.57	4.22±3.58	84.73±6.60	0.87±0.38	4.03±3.34
UNet [26]	79.74±8.27	1.14±0.91	6.55±5.10	81.60±7.99	1.05±0.60	6.16±5.07
UNet++ [23]	79.68±8.17	1.17±0.95	6.91±5.50	81.63±7.93	1.05±0.62	7.10±5.46

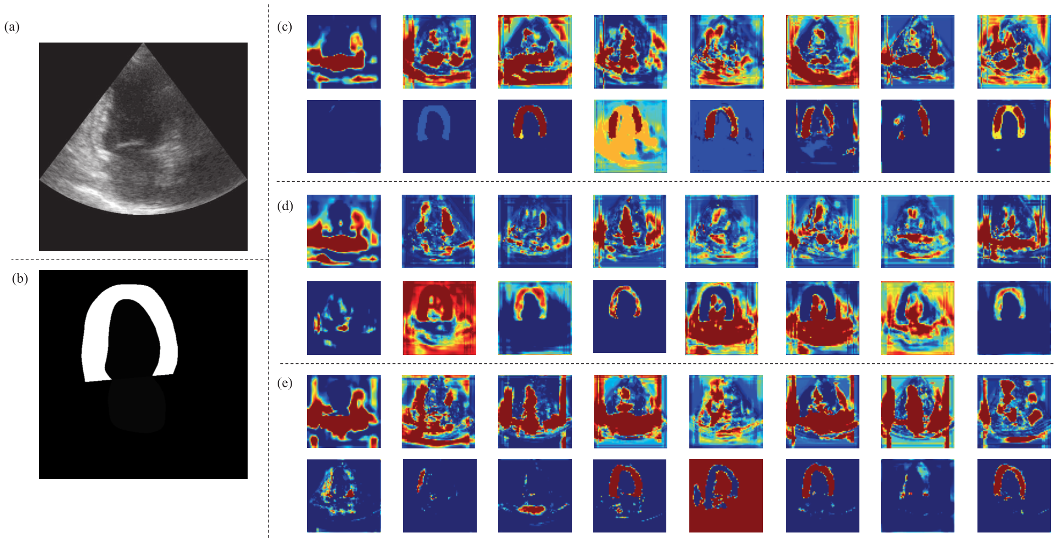


Fig. 8. FiLM design achieved more effective feature enhancement than other settings. (a) Input image. (b) Corresponding label. The performance of the three settings in Table I is illustrated in (c)–(e), respectively.

removing the small area on the segmentation result. Results in Table III showed that the proposed method achieved for most of the metrics compared with other methods. Some methods, such as PLANet and ACNN in our experiments, have been integrated with the label coherence information and shape prior to the learning of anatomical structures. Notably, based on an encoder–decoder design, CPFNet outperformed the other methods and performed close to our method, demonstrating that global/multiscale information fusion on context information can also achieve better segmentation performance. Our method applied the ground truth to capture the anatomical information indirectly can achieve higher performance. Furthermore, we performed a statistical comparison of the Dice results using paired t-test with a confidence interval of 0.95. MCAL is compared to CPFNet for statistical significance, and the p values specified to the Dice of A4C/ES/ED are 0.004100/0.000178/0.001100. It can be seen that the proposed method significantly outperforms CPFNet with $p < 0.05$.

E. Ablation Study

We investigated the contributions of each module of our method to the segmentation performance by different configurations in the public CAMUS. We also explored the impact of hyperparameters in the loss function on the behavior of MCAL by grid-searching. In the hyperparameter setting experiments, the patients with annotations in CAMUS were randomly divided into training (410) and evaluation (40) datasets. In the ablation study experiments, we applied the same training strategy with the comparison study. All the experiments were conducted under the same training and evaluation methods as in [37]. We determined the model with the best performance for each group on the Dice coefficient.

TABLE IV

ABLATION RESULTS FOR MCAL WITH DIFFERENT SETTINGS ON CAMUS. BOLD NUMBERS REPRESENT THE BEST RESULTS OBTAINED

Spatial factor	FiLM	BDTW	A2C			A4C		
			Dice(%)	d_m(mm)	d_H(mm)	Dice(%)	d_m(mm)	d_H(mm)
✓	✓	✓	**85.33±5.65**	**0.92±0.37**	**3.54±2.45**	**85.85±5.59**	**0.73±0.30**	**3.61±2.98**
C	✓	✓	84.11±6.73	0.97±0.50	4.82±4.18	84.51±6.25	0.76±0.30	3.98±3.74
✓	○	✓	83.83±6.59	0.99±0.50	4.80±3.74	83.30±6.98	0.83±0.32	4.19±3.37
✓	×	×	84.19±6.64	0.97±0.46	4.46±3.43	84.48±6.33	0.77±0.32	3.94±3.82
✓	✓	×	84.31±6.54	0.96±0.45	4.45±3.46	84.39±6.24	0.78±0.35	3.92±3.79
✓	○	×	84.22±6.85	0.97±0.45	4.62±3.74	84.40±6.47	0.77±0.29	3.98±3.63
C	✓	×	83.89±6.56	0.98±0.47	4.73±3.77	83.33±6.84	0.82±0.33	4.17±3.67
✓	×	✓	84.07±6.57	0.97±0.46	4.72±3.92	84.14±6.49	0.78±0.32	3.87±3.39
×	×	✓	84.07±6.64	0.98±0.46	4.47±3.49	83.87±7.34	0.81±0.45	3.93±3.30
×	×	×	83.69±7.05	1.01±0.59	4.45±3.41	84.02±6.93	0.79±0.32	3.69±2.74
			ED			ES		
✓	✓	✓	**85.10±5.58**	**0.83±0.36**	**3.42±2.28**	**86.08±5.59**	**0.81±0.38**	**4.01±3.43**
C	✓	✓	83.84±6.42	0.86±0.48	4.28±3.66	84.78±6.54	0.87±0.36	4.53±4.27
✓	○	✓	82.94±6.72	0.90±0.46	4.53±3.47	84.20±6.80	0.91±0.39	4.46±3.67
✓	×	×	83.70±6.30	0.88±0.46	4.40±3.69	84.80±6.79	0.88±0.39	4.22±3.88
✓	✓	×	83.87±6.14	0.87±0.42	4.17±3.38	84.83±6.43	0.88±0.40	4.17±3.78
✓	○	×	83.75±6.66	0.87±0.43	4.39±3.68	84.88±6.61	0.87±0.35	4.20±3.71
C	✓	×	82.96±6.58	0.90±0.43	4.57±3.70	84.26±6.76	0.90±0.39	4.33±3.74
✓	×	✓	83.51±6.42	0.87±0.43	4.36±3.59	84.70±6.58	0.88±0.38	4.24±3.78
×	×	✓	83.43±6.63	0.88±0.41	4.29±3.33	84.50±7.32	0.91±0.51	4.12±3.47
×	×	×	83.32±6.88	0.90±0.55	4.08±2.97	84.38±7.06	0.91±0.42	4.06±3.26

TABLE V

COMPARISON OF EXISTING METHODS FOR LV SEGMENTATION FOR CLINICAL DEPLOYMENT

Methods	Models	Description	Clinical Limitation
Non-deep learning	ACM	Minimizing an energy function under the influence of different forces and constraints	Require user-imposed guidance to achieve high accuracy
	AAM	Describing the image appearance and the shape as a statistical shape-appearance model	Require consistent shape prior over a large database
Deep learning	UNet	Encoder-decoder	Poor model generalization; Limit performance on the LV segmentation
	UNet++	Highly flexible feature fusion	
	HarDNet-MSE	Low memory traffic backbone	
	PLANet	Features enhancement by label coherence learning	Complicate computation; Lack of temporal information
	CPFNet	Feature enhancement by preserving abstract spatial information	Lack of temporal information; Poor model generalization; Lack of model interpretability
	SegNet	Pooling indices are applied in the max-pooling step	
	CNN	Labels are also used as anatomical prior	
	Ours	Labels are also used for feature enhancement	

1) Ablation for Spatial Factor and FiLM: To verify the effectiveness of FiLM design on the relative segmented features, we replaced the rescale and offset coefficients in FiLM design by concatenating the dimensional expanding on samples from learned spatial factors z_s with successive convolution layers, which is represented by 'C' in the spatial factor column. Alternatively, the rescale and offset coefficients are derived defectively from the spatial space s through successive convolution operations, which is represented by "○" in the FiLM column. The results are shown in Table IV. We observed a performance drop when the rescale and offset designs are replaced with successive convolution layers, indicating the effectiveness of the FiLM structure on the feature enhancement. Similarly, deriving the rescale and offset directly from the spatial space performed worse, demonstrating the necessity of the FiLM structure. The visualization results in Fig. 8(c) and (e) are evidence that the FiLM design performs better than the concatenation operation. It demonstrates that an affine transformation to each channel of the feature map helps to enhance the segmented-relevant features. However, the learned spatial factor is more effective for guiding feature enhancement [see Fig. 8(d)]. It is obvious that the performance has been greatly improved on the condition that the scale and offset parameters are derived from the learned spatial factors.

2) Ablation for BDTW: Based on these configurations, we investigated the performance of BDTW on segmentation. Results in Table IV displayed that improvement is limited when adding the BDTW to the baseline. However, the Dice dropped slightly when adding the BDTW to the spatial factor, while the distance error metrics improved, because it is difficult to balance the spatial distribution and the boundary region under two different scales without intermediate operation in the training stage. An improvement was illustrated in Table IV when the FiLM design was added to the configurations. In conclusion, the joint of FiLM structure on the spatial factors and BDTW has improved the performance of segmentation.

V. Discussion

This work tackles the challenge of myocardium segmentation because of the fuzzy boundary caused by the modality imaging characteristic. This segmentation task can be solved by deep neural networks based on different settings. Since the ground-truth label as class associations on segmentation lacks effective feature enhancement of segmented region, we leverage anatomical knowledge learned through ground-truth labels to infer and strengthen the segmented parts. By regulating the distribution discrepancy of two posterior probabilities, which are approximated from the sampled input and labels, we can modulate the learned anatomical feature maps based on the FiLM. Furthermore, we apply BDTW to assign a higher weight to boundary region pixels in each training batch. Hence, the proposed model can achieve high performance on both segmentations.

This work has limitations. The temporal relevance caused by cardiac movement is neglected in this article. Studies on 2-D echocardiography segmentation have been proven to be effective by adding temporal information [27], [42]. This makes sense as we currently studied anatomical knowledge through labels to infer segmented parts and discriminate boundary pixels. However, the segmentation accuracy is limited. Our method still needs to be improved before the clinical deployments. Because we just train and test our model based on data from two different devices, the model generalization has not been fully verified. In the future, it might be expected to embed a temporal correlation of the successive frames or the model generalization to get a better result (see Table V).

In general, the proposed method of myocardium segmentation can be applied to other image modalities when we analyze medical images by segmentation. For instance, it can be used for anatomical segmentation by retraining the network based on new images and labels.

VI. Conclusion

Accurate LV segmentation in 2-D echocardiography is significant for cardiovascular disease diagnosis and assessment of cardiac function. However, it is difficult to discriminate between the myocardium and chamber due to the characteristic of echos. A new method named MCAL, which makes use of prior anatomical information and constraint of the predicted map, is proposed to infer the segmented structure and discriminate the boundary pixels. We apply a KL divergence to learn a spatial factor of the raw input that can account for segmentation. Furthermore, the spatial factor modulates the encoded spatial space by FiLM to strengthen the critical segmented structure information in relative channels. A skip convolution fusion block combines the semantical information from the encoder with the relatively rich anatomical information in the decoder and solves the segmentation by a bottom-up structure. In addition, we introduce the BDTW to weight the binary cross-entropy loss to force the network to focus on the border neighborhood pixels in each training epoch. Finally, we test the proposed method on two different datasets, and experiment results reveal that the proposed method can improve the myocardial segmentation performance.

References

[1] R. M. Lang *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.-Cardiovascular Imag.*, vol. 16, no. 3, pp. 233–271, 2015.

[2] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, Jun. 2001.

[3] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *Int. J. Comput. Vis.*, vol. 1, no. 4, pp. 321–331, Jan. 1988.

[4] M. Rousson and N. Paragios, "Shape priors for level set representations," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2002, pp. 78–92.

[5] P. Tang, X. Yang, Y. Nan, S. Xiang, and Q. Liang, "Feature pyramid nonlocal network with transform modal ensemble learning for breast tumor segmentation in ultrasound images," *IEEE Trans. Ultrason., Ferroelectr., Freq. Control*, vol. 68, no. 12, pp. 3549–3559, Dec. 2021.

[6] Y. Fang, H. Huang, W. Yang, X. Xu, W. Jiang, and X. Lai, "Nonlocal convolutional block attention module V Net for gliomas automatic segmentation," *Int. J. Imag. Syst. Technol.*, pp. 1–16, Jul. 2021, doi: 10.1002/ima.22639.

[7] H. Ravishankar, R. Venkataramani, S. Thiruvenkadam, P. Sudhakar, and V. Vaidya, "Learning and incorporating shape models for semantic segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Quebec City, QC, Canada: Springer, 2017, pp. 203–211.

[8] M. Tofighi, T. Guo, J. K. P. Vanamala, and V. Monga, "Deep networks with shape priors for nucleus detection," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 719–723.

[9] O. Oktay *et al.*, "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.

[10] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction," in *Proc. Int. Conf. Artif. Neural Netw.* Berlin, Germany: Springer, 2011, pp. 52–59.

[11] R. Ge *et al.*, "*K*-Net: Integrate left ventricle segmentation and direct quantification of paired echo sequence," *IEEE Trans. Med. Imag.*, vol. 39, no. 5, pp. 1690–1702, May 2020.

[12] R. Ge *et al.*, "PV-LVNet: Direct left ventricle multitype indices estimation from 2D echocardiograms of paired apical views with deep neural networks," *Med. Image Anal.*, vol. 58, Dec. 2019, Art. no. 101554.

[13] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32, no. 1, pp. 1–10.

[14] J. M. Dias and J. M. Leitao, "Wall position and thickness estimation from sequences of echocardiographic images," *IEEE Trans. Med. Imag.*, vol. 15, no. 1, pp. 25–38, Feb. 1996.

[15] V. Chalana, D. T. Linker, D. R. Haynor, and Y. Kim, "A multiple active contour model for cardiac boundary detection on echocardiographic sequences," *IEEE Trans. Med. Imag.*, vol. 15, no. 3, pp. 290–298, Jun. 1996.

[16] Z. Tao and H. Tagare, "Tunneling descent for m.a.p. active contours in ultrasound segmentation," *Med. Image Anal.*, vol. 11, no. 3, pp. 266–281, Jun. 2007.

[17] M. Mignotte, J. Meunier, and J.-C. Tardif, "Endocardial boundary e timation and tracking in echocardiographic images using deformable template and Markov random fields," *Pattern Anal. Appl.*, vol. 4, no. 4, pp. 256–271, Nov. 2001.

[18] T. Dietenbeck, M. Alessandrini, D. Barbosa, J. D'hooge, D. Friboulet, and O. Bernard, "Detection of the whole myocardium in 2D-echocardiography for multiple orientations using a geometrically constrained level-set," *Med. Image Anal.*, vol. 16, no. 2, pp. 386–401, 2012.

[19] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017.

[20] S. Liu *et al.*, "Deep learning in medical ultrasound analysis: A review," *Engineering*, vol. 5, no. 2, pp. 261–275, 2019.

[21] 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.

[22] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes," *IEEE Trans. Med. Imag.*, vol. 37, no. 12, pp. 2663–2674, Dec. 2018.

[23] Z. Zhou, M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," *arXiv:1807.10165*, 2018.

[24] 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. 2017.

[25] D. Fourure, R. Emonet, E. Fromont, D. Muselet, A. Tremeau, and C. Wolf, "Residual conv-deconv grid network for semantic segmentation," 2017, *arXiv:1707.07958*.

[26] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Munich, Germany: Springer, 2015, pp. 234–241.

[27] S. Leclerc *et al.*, "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.

[28] D. Ouyang *et al.*, "Video-based AI for beat-to-beat assessment of cardiac function," *Nature*, vol. 580, pp. 252–256, 2020, doi: 10.1038/s41586-020-2145-8.

[29] M. H. Jafari *et al.*, "A unified framework integrating recurrent fully-convolutional networks and optical flow for segmentation of the left ventricle in echocardiography data," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Granada, Spain: Springer, 2018, pp. 29–37.

[30] M. Li *et al.*, "Recurrent aggregation learning for multi-view echocardiographic sequences segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Shenzhen, China: Springer, 2019, pp. 678–686.

[31] Q. Yue, X. Luo, Q. Ye, L. Xu, and X. Zhuang, "Cardiac segmentation from LGE MRI using deep neural network incorporating shape and spatial priors," 2019, *arXiv:1906.07347*.

[32] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*.

[33] C. Ouyang, K. Kamnitsas, C. Biffi, J. Duan, and D. Rueckert, "Data efficient unsupervised domain adaptation for cross-modality image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Shenzhen, China: Springer, 2019, pp. 669–677.

[34] F. Wu and X. Zhuang, "Unsupervised domain adaptation with variational approximation for cardiac segmentation," *IEEE Trans. Med. Imag.*, vol. 40, no. 12, pp. 3555–3567, Dec. 2021.

[35] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[36] M. Berman, A. R. Triki, and M. B. Blaschko, "The Lovász–Softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4413–4421.

[37] F. Liu, K. Wang, D. Liu, X. Yang, and J. Tian, "Deep pyramid local attention neural network for cardiac structure segmentation in two-dimensional echocardiography," *Med. Image Anal.*, vol. 67, Jan. 2021, Art. no. 101873.

[38] W. R. Crum, O. Camara, and D. L. G. Hill, "Generalized overlap measures for evaluation and validation in medical image analysis," *IEEE Trans. Med. Imag.*, vol. 25, no. 11, pp. 1451–1461, Nov. 2006.

[39] S. Feng *et al.*, "CPFNet: Context pyramid fusion network for medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 10, pp. 3008–3018, Oct. 2020.

[40] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[41] C.-H. Huang, H.-Y. Wu, and Y.-L. Lin, "HarDNet-MSEG: A simple encoder-decoder polyp segmentation neural network that achieves over 0.9 mean dice and 86 FPS," 2021, *arXiv:2101.07172*.

[42] H. Wei *et al.*, "Temporal-consistent segmentation of echocardiography with co-learning from appearance and shape," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Lima, Peru: Springer, 2020, pp. 623–632.

Xiaoxiao Cui is currently pursuing the Ph.D. degree with the School of Information Science and Engineering, Shandong University, Jinan, China.

Her current research interest is medical image analysis.

Pengfei Zhang received the Ph.D. degree in medicine from Shandong University, Jinan, China, in 2005.

He is currently a Doctor of Medicine and a Professor of Shandong University. He is also the Chief Physician with the Qilu Hospital, Shandong University. In conformal cardiac ultrasound architecture and its supporting artificial intelligence chips, beam synthesis, ultrasonic image intelligent analysis, ultrasonic conformal material preparation technology, and ultrasound engineering and clinical applications, he has published more than 50 SCI papers. He has applied for 18 invention patents and six utility model patents. He has been granted six patents and three PCT patents.

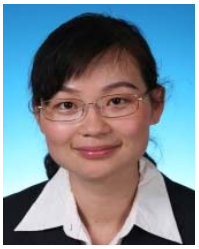
Dr. Zhang received the 11th The World Federation of Ultrasound Medicine Award and the 20th Annual Meeting of the American Society of Echocardiography, Characteristic Original Research.

Yujun Li received the Ph.D. degree from the Harbin Institute of Technology, Harbin, China, in 2001.

He is currently a Full Professor with the Department of Information Science and Engineering, Shandong University, Jinan, China. His research interests include deep learning, natural language processing, and sentiment analysis.

Zhi Liu received the Ph.D. degree from the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China, in 2008.

He is currently an Associate Professor with the School of Information Science and Engineering, Shandong University, Jinan, China, where he is the Head of the Intelligent Information Processing Group. His current research interests are in applications of computational intelligence to linked multicomponent big data systems, medical images in the neurosciences, multimodal human–computer interaction, remote sensing image processing, content-based image retrieval, semantic modeling, data processing, classification, and data mining.



Xiaoyan Xiao received the Ph.D. degree from Shandong University, Jinan, China, in 2010.

She is currently a Visiting Scholar with the Harvard Medical School, Boston, MA, USA, and an Attending Physician with the Qilu Hospital, Shandong University.

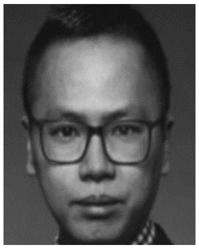
Yang Zhang received the Medical degree from Shandong University, Jinan, China, in 2010.

She was a Visiting Scholar with Cornell University, Ithaca, NY, USA, from 2014 to 2015. She is currently an Associate Chief Physician with the Department of Radiology, Qilu Hospital, Shandong University. She specializes in neuroimaging and cardiovascular imaging. She runs two active programs of Neuro-MR research at Shandong University.

Longkun Sun is currently pursuing the Medical master's degree in cardiology with the School of Clinical Medicine, Shandong University, Jinan, China.

He is studying at the Qilu Hospital, Shandong University. His current research interest is echocardiography.

Lizhen Cui (Member, IEEE) is currently a Full Professor and a Doctoral Supervisor with Shandong University, Jinan, China. He is also a Visiting Scholar with the Georgia Institute of Technology, Atlanta, GA, USA, and a Visiting Professor with Nanyang Technological University, Singapore. He is also appointed as the Dean of the Undergraduate School, the Dean and the Deputy Party Secretary of the School of Software, the Co-Director of the Joint SDU-NTU Centre for Artificial Intelligence Research (C-FAIR), the Director of the Research Center of Software and Data Engineering, and the Associate Director of the National Engineering Laboratory for E-Commerce Transaction Technologies, Shandong University. His research interests include data intelligence and cognitive informatics, and trusted artificial intelligence.



Guang Yang (Senior Member, IEEE) received the M.Sc. degree in vision imaging and virtual environments from the Department of Computer Science, University College London, London, U.K., in 2006, and the Ph.D. degree in medical image analysis from the Centre for Medical Image Computing (CMIC), Department of Computer Science and Medical Physics, University College London, in 2012.

He is currently an Honorary Lecturer with the Neuroscience Research Centre, Cardiovascular and Cell Sciences Institute, St. George's, University of London, London. He is also an Image Processing Physicist and an Honorary Senior Research Fellow with the Cardiovascular Research Centre, Royal Brompton Hospital, Chelsea, London. He is also with the National Heart and Lung Institute, Imperial College London, London.

Dr. Yang is also a member of the International Society for Magnetic Resonance in Medicine (ISMRM) and the Society of Photo-Optical Instrumentation Engineers (SPIE). (Based on a document published on August 17, 2020.)

Shuo Li (Senior Member, IEEE) has been the Founder and the Scientific Director of the Digital Imaging Group of London (DIG), London, ON, Canada, since 2007. He has been a Scientist with the Lawson Health Research Institute, London, ON, Canada, since 2015. He is currently an Associate Professor with Western University, London, where he is also an Academic Artificial Intelligence Scientist with a background in both medical imaging and computer science and software engineering. As a primary supervisor, he has mentored over 60 trainees, including postdoctoral fellows, Ph.D. students, M.Sc. students, undergraduates, and visiting scholars. His current research interests include the development of artificial intelligence systems solving the most challenging clinical and fundamental data analytics problems in radiology, urology, surgery, rehabilitation, and cancer, with an emphasis on the innovations of our learning schemes (regression learning, deep learning sparse learning, spectral learning, and manifold learning).

Dr. Li also serves as a Board Member of the Medical Image Computing and Computer-Assisted Intervention Society (MICCAI) and the General Chair of the MICCAI Conference 2022. He has been frequently invited to review grant applications for several international funding agencies. He has been actively serving multiple professional societies in various leadership capacities.