IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 5, 2020, accepted September 17, 2020, date of publication September 23, 2020,
date of current version October 2, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3026073

Bridge Segmentation Performance Gap via Evolving Shape Prior

CHAOYU CHEN[1,2], XIN YANG[1,2], HAORAN DOU[1,2], RUOBING HUANG[1,2], XIAOQIONG HUANG[1,2], XU WANG[1,2], CHONG DUAN[3], SHENGLI LI[4], WUFENG XUE[1,2], (Member, IEEE), PHENG ANN HENG[5], (Senior Member, IEEE), AND DONG NI[1,2], (Member, IEEE)

[1]National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, Health Science Center, School of Biomedical Engineering, Shenzhen University, Shenzhen 518060, China
[2]Medical Ultrasound Image Computing (MUSIC) Laboratory, Shenzhen University, Shenzhen 518060, China
[3]Department of Early Clinical Development, Pfizer Inc., Cambridge, MA 02130, USA
[4]Department of Ultrasound, Affiliated Shenzhen Maternal and Child Healthcare Hospital, Nanfang Medical University, Guangzhou, China
[5]Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong

Corresponding author: Dong Ni (nidong@szu.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2019YFC0118300, in part by the Shenzhen Peacock Plan under Grant KQTD2016053112051497 and Grant KQJSCX20180328095606003, in part by the Natural Science Foundation of China under Grant 61801296, and in part by the Shenzhen Basic Research under Grant JCYJ20190808115419619.

ABSTRACT Deep neural networks are very compelling for medical image segmentation. However, deep models often suffer from notable performance drops in real clinical settings due to the complex appearance shift in daily scannings. Domain adaptation partially addresses the problem between imaging domains. However, it heavily depends on the expensive re-collection and re-training for domain-specific datasets and thus is not applicable to domain-agnostic images. In this paper, we propose a *case adaptation* strategy aiming to bridge the segmentation performance gap on domain-agnostic images. Our contribution is three-fold. First, we design a general self-supervised learning framework for *case adaptation*, which exploits its predictions as supervision to drive the adaptation. Without extra annotations and any burden on model complexity, the framework enables trained deep models at-hand to directly segment domain-agnostic testing images. Second, we propose a novel Evolving Shape Prior (ESP) which recursively introduces strong shape knowledge into networks and evolves with the adaptation procedure to provide adaptive supervision. ESP can stabilize self-supervised learning and guide it to move towards model convergence. Third, we perform extensive experiments on 10 datasets with different levels of difficulty and typical appearance shifts blended, proving our framework is a promising solution in reducing segmentation performance degradation. Through this work, we investigate the feasibility of *case adaptation* as a general strategy in enhancing the robustness of deep segmentation networks, with comprehensive analyses proving its efficacy and efficiency.

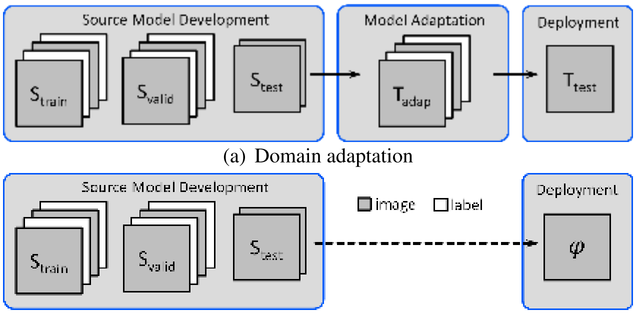
INDEX TERMS Deep neural networks, medical image segmentation, case adaptation, self-supervision, evolving shape prior.

I. INTRODUCTION

The great resurgence of deep learning brings profound and lasting impact on medical image segmentation [1], [2]. However, due to the data dependency and lack of generalization ability, deep segmentation models often lose their power in practical scenarios, especially in daily clinical settings [3], [4]. As shown in the left of Fig. 1(a), high accuracies achieved by deep learning models are often reported within a pre-defined domain $\mathcal{S}$, where training, validation and testing images share a coherent appearance distribution. However, what has been required in the real clinical settings is that, the model should be independent of its source domain $\mathcal{S}$ and work consistently on each domain-agnostic testing image $\boldsymbol{\varphi}$ (Fig. 1(b)) which presents unpredictable appearance shift (Fig. 2). The segmentation performance gap between the model development phase on $\mathcal{S}$ and the deployment phase on $\boldsymbol{\varphi}$ can be large and has been recognized in some recent studies [4]–[6].

Domain adaptation [5] is an alternative approach to deal with this gap (Fig. 1(a)). It usually unifies model input or modulating the model itself for a new domain $\mathcal{T}$ with

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



(a) Domain adaptation

(b) Case adaptation

FIGURE 1. Schematic view to show the difference between (a) domain adaptation and (b) case adaptation. $\mathcal{S}$ and $\mathcal{T}$ stand for two independent domains. Testing image φ in model deployment phase is domain-agnostic.

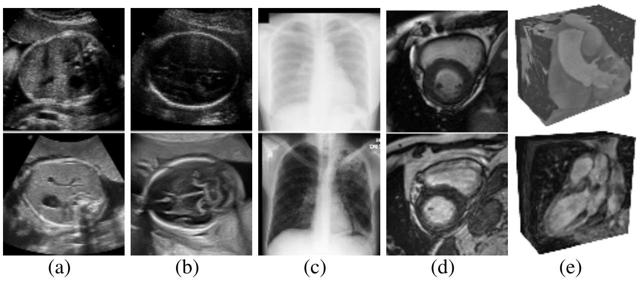


FIGURE 2. Illustration of appearance shift in medical image segmentation. Top and bottom rows: (a) ultrasound images of fetal abdomen from different scanners, (b) ultrasound images of fetal head from different timepoints, (c) X-ray images of chest from different protocols, (d) cine-MR images of heart from different clinical centres (e) CT and MR volumes of whole heart. Similar object shape is shared across datasets in each column.

the assistance from a labeled or unlabeled dataset $\mathcal{T}_{adap}$. Drozdzal *et al.* [7] proposed a lightweight Fully Convolutional Network to learn to normalize the medical image appearance. However, the learned normalization is only effective for small appearance variations. Generative adversarial network (GAN) [8] can generate realistic style translation between two distinctive medical image domains [9] and thus enable the appearance harmonization. To further enhance the boundary sharpness in CT-MR translation results, segmentation based shape consistency loss was proposed in cycled GAN [10]–[12]. By aligning features of different domains, Kamnitsas *et al.* [13] exploited an adversarial scheme to teach the network to learn source-invariant representations for brain lesion segmentation on images from different scanners. A similar idea also appeared in [14] for appearance-invariant breast cancer classification in histopathology images.

Despite the effectiveness of domain adaptation, it still has two critical drawbacks in clinical scenarios. First, it only extends models in $\mathcal{S}$ to a fixed and pre-defined domain $\mathcal{T}$ (Fig. 1(a)). As shown in Fig. 2, there are many imaging factors in daily scanning for the same examination, such as different scanners, operators, protocols, timepoints, etc. These blended factors make every testing image present a unique appearance shift and can be domain-blinded against segmentation models. Thus, it is infeasible to clearly define a bounded domain to run domain adaptation [3]. Second, domain adaptation greatly depends on $\mathcal{T}_{adap}$. However, for each subject, only a few images (maybe just one) would be acquired for each task [6]. Collecting a large amount of images and labels on-site to build a $\mathcal{T}_{adap}$ and then modulating the pre-trained models are impractical. Therefore, as we propose, *case adaptation* (Fig. 1(b)) could be a better strategy to fulfill the requirements of clinical settings. This new strategy discards $\mathcal{T}_{adap}$ and focuses on directly deploying the trained deep models on each domain-agnostic image.

Recently, efforts have been devoted to reducing the dependency on $\mathcal{T}_{adap}$. Gibson *et al.* [4] proved that utilizing annotated images of as few as 8 subjects from the unseen site for calibration is possible to address the inter-site prostate segmentation in MR images. In [15], the model fine-tuned with a single annotated image achieved comparable results against full dataset trained competitors for lesion segmentation. However, the choice of annotated images may have a considerable impact on performances. Wang *et al.* [16] relaxed the annotation requirement by proposing an interactive scribble based fine-tuning on images. Huang *et al.* [17] made further attempts to re-train student models with pseudo labels generated by teacher models trained under a data- and model-distillation scheme. In [6], for prenatal ultrasound image segmentation, Yang *et al.* inherited the fine-tuning fashion. They utilized a network-based renderer to unify input appearance, and an adversarial structure loss to exempt the need for extra annotations. These studies suggest that, despite the slight computation cost, case adaptation, i.e., image-specific fine-tuning, is promising to bridge segmentation performance gaps on domain-agnostic images.

Using label proxy generated by the model itself, such as predictions, to supervise and fine-tune the model, *self-supervised learning* sheds light on bypassing manual annotations for fine-tuning [18], [19]. However, solely resorting to its prediction, vanilla self-supervised learning is insufficient in combating segmentation performance gaps. Accurate and informative label proxy is crucial for stable self-supervised learning and the model convergence.

Model-based methods, i.e. atlas-based model [20], [21], statistical shape model [22], [23] leverage the shape prior to guide the model segmentation. For instance, Cheng *et al.* [24] introduced an active contour framework with shape prior for the vessel segmentation. Although effective, these model-based methods are usually task-specific or modality-specific. Zheng *et al.* [25] investigated the effectiveness of shape priors learned from a different modality to improve the segmentation accuracy on the target modality. Inspired by [6], [25], we find that shape prior is a strong knowledge across different imaging conditions (Fig. 2) and is beneficial to self-supervised learning. Nevertheless, the traditional shape prior is often modeled with handcrafted features and limited deformation ranges [25], [26], and thus can not be easily incorporated into the dynamic self-supervised learning process.

In this paper, we try to cope with the segmentation performance gap in real clinical settings by following the proposed *case adaptation*. Our contribution is three-fold.

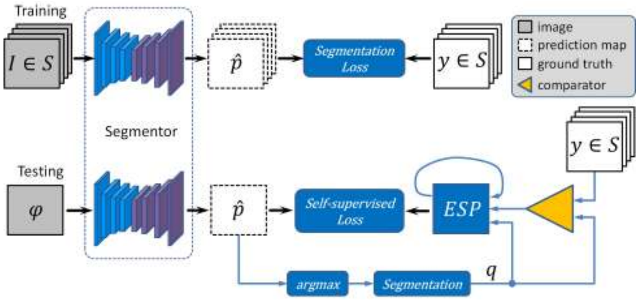


FIGURE 3. Schematic view of our proposed framework. Segmentor is firstly trained following a classic pipeline. Our recursive self-supervised learning scheme then promptly tunes the segmentor to directly segment the domain-agnostic testing image φ without any manual assistance.

First, we design a general self-supervised learning-based scheme, which is annotation-free and explores its prediction for fine-tuning. It preserves model complexity and enables trained models to directly segment domain-agnostic testing images. Second, we propose a novel Evolving Shape Prior (ESP) which recursively introduces strong shape knowledge into networks and evolves with the fine-tuning procedure to provide adaptive supervision. ESP proves to be very effective in making self-supervised learning run stably and guiding the fine-tuning converge towards a preferred target. Third, under extensive experiments on 10 datasets with different levels of difficulty and typical appearance shifts, our framework proves to be a promising solution in bridging various segmentation performance gaps. These datasets include images from different scanners (ultrasound of fetal abdomen, Fig. 2(a)), timepoints (ultrasound of fetal head, Fig. 2(b)), protocols (X-ray of lung, Fig. 2(c)), clinical centres (cine-MR of heart, Fig. 2(d)) and modalities (CT and MR of whole heart, Fig. 2(e)). As a by-product, our framework also presents the potentials to be a simple and efficient refinement strategy to improve segmentation on source domains. Our proposed method is general and easy to implement to enhance the robustness of deep segmentation networks.

II. METHODOLOGY

Fig. 3 depicts our proposed framework. Top row shows a typical pipeline to train a segmentor with the loss calculated between prediction maps and ground truth labels. A certain amount of paired training images $\boldsymbol{I}$ and pixel-wise labels $\boldsymbol{y}$ from domain $\mathcal{S}$ are required. Bottom row is our proposed self-supervised learning scheme with ESP for testing. Trained segmentor deploys on the domain-agnostic testing image $\boldsymbol{\varphi}$ and generates prediction maps $\hat{\boldsymbol{p}}$. Segmentation $\boldsymbol{q}$ is then produced by an *argmax* layer and goes to combine similar label candidates in $\boldsymbol{y}$ to recursively stimulate the evolution of ESP. In a few iterations supervised by ESP, the segmentor trained in $\mathcal{S}$ can quickly and stably learn to tackle the appearance discrepancy in domain-agnostic image $\boldsymbol{\varphi}$ and generate satisfying segmentation to minimize the self-supervised loss.

A. BACKBONE OF THE SEGMENTATION NETWORK

Our proposed framework is not exclusively designed for a specific segmentation network. In this paper, as shown in Fig. 4, we choose a classic and powerful U-net [27] like network coupled with hybrid loss functions and a deep supervision scheme as a backbone to achieve competitive performances.

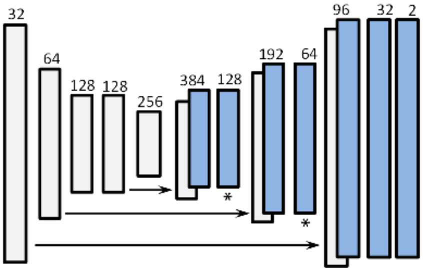


FIGURE 4. Architecture of segmentation network. Digits denote the number of feature maps. Black stars denote the anchor point of auxiliary loss functions.

As datasets depicted in Fig. 2, our segmentation network needs to overcome several challenges: (i) boundary ambiguity caused by noise and low contrast, (ii) boundary deficiency caused by occlusion and signal dropout, (iii) class-imbalance among different classes. Facing these challenges, we use a hybrid loss function ($\mathcal{L}_{hyb}$, Eq. 1) as proposed in [28], which combines weighted cross-entropy loss ($\mathcal{L}_{wCross}$) and multi-class Dice Similarity Coefficient based loss ($\mathcal{L}_{mDSC}$). Both loss components are helpful against class-imbalance. The former one is preferred in preserving boundary details and the latter emphasizes global shape similarity to generate compact segmentation. A hybrid version of these two losses performs better than each alone [28].

Gradient vanishing often adversely affects the learning of deep networks. Deep supervision mechanism can relieve the problem by exposing shallow layers to the extra and composite supervision of d auxiliary loss functions via d side paths. we refer readers to [29] for more details. In this paper, we attached two auxiliary loss functions in our network ($d = 2$), denoted as * in Fig. 4. All loss functions in our network are in a hybrid version. Final loss function $\mathcal{L}_{seg}$ is thus formulated as Eq. 2,

$$\mathcal{L}_{hyb} = \mathcal{L}_{wCross} + \lambda \mathcal{L}_{mDSC} \tag{1}$$

$$\mathcal{L}_{seg} = \mathcal{L}^{0}_{hyb} + \sum_{i=1}^{d} \eta_i \mathcal{L}^{i}_{hyb} \tag{2}$$

where λ is set as 100 to balance the scale of $\mathcal{L}_{wCross}$ and $\mathcal{L}_{mDSC}$. $\mathcal{L}^{0}_{hyb}$ is the main loss function, while $\mathcal{L}^{i}_{hyb}$ $(i \geq 1)$ is the auxiliary loss. η_i is the weight of $\mathcal{L}^{i}_{hyb}$ in final loss. The auxiliary loss in shallow layer is assigned with smaller weight ($\eta_1 = 0.4$) than that in deeper layer ($\eta_2 = 0.8$) to avoid the network excessively focusing on the boundary details and ignoring the semantic representations in the shallow layer [30].

Input images to our network are normalized as zero mean and unit variance by itself. Small convolution kernel with size of 3 is utilized in all convolutional layers (2D kernel for 2D segmentation, Fig. 2(a)(b)(c)(d). 3D kernel for volumetric segmentation, Fig. 2(e)). Each convolutional layer is followed by a batch normalization layer and a rectified linear unit. For the four diverse tasks as shown in Fig. 2, our network design proves to be general and effective in achieving satisfying segmentation in each source domain.

B. SELF-SUPERVISED LEARNING SCHEME

It is hard for well-trained segmentors to generalize to unseen images with appearance shifts. Fully-supervised fine-tuning tackles the problem with many annotations [4], [15]. Weakly-supervised fine-tuning for a specific image reduces the need of extra annotations [6], [16], [17]. These studies provide a strong hint that, under a weakly- or un-supervised setting, bridging the segmentation performance gap with case adaptation can be feasible.

Self-supervised learning (*SSL*) scheme occurs in our view as an attractive solution for case adaptation. Supervising a learning process with the label proxy generated by the model itself and thus being annotation-free is the core idea of the SSL scheme. After pre-trained on a source domain, the model in the scheme needs to learn to simultaneously update itself and also the proxy to achieve reasonable segmentation on unseen images. As shown in Fig. 3, when applying the SSL to our case adaptation based segmentation, the problem can be formulated as follows:

$$\min_{\boldsymbol{w},\hat{\boldsymbol{y}}_\varphi} \mathcal{L}_F(\boldsymbol{w},\hat{\boldsymbol{y}}_\varphi) = \sum_{s=1}^{S} \mathcal{L}_{\mathcal{S}}(\boldsymbol{y}_s,\hat{\boldsymbol{p}}(\boldsymbol{w},\boldsymbol{I}_s)) + \mathcal{L}_\varphi(\hat{\boldsymbol{y}}_\varphi,\hat{\boldsymbol{p}}(\boldsymbol{w},\boldsymbol{\varphi})) \quad (3)$$

where $\boldsymbol{I}_s$ is the image in source domain $\mathcal{S}$ ($s = 1, 2, \ldots, S$), $\boldsymbol{y}_s$ is the pixel-wise label of $\boldsymbol{I}_s$. $\boldsymbol{w}$ is learnable weight of network, $\hat{\boldsymbol{p}}$ is the network predictions representing class probabilities. $\mathcal{L}_{\mathcal{S}}$ is the loss in pre-training network and set as $\mathcal{L}_{seg}$ in Eq. 2. $\mathcal{L}_{\mathcal{S}}$ is minimized by optimizing $\boldsymbol{w}$ on $\mathcal{S}$. $\boldsymbol{\varphi}$ is domain-agnostic testing image. $\hat{\boldsymbol{y}}_\varphi$ is the proxy to supervise the case adaptation, which is generated by network itself and then optimized during case adaptation. $\mathcal{L}_\varphi$ is self-supervised loss for case adaptation. Both $\boldsymbol{w}$ and $\hat{\boldsymbol{y}}_\varphi$ are tuned only on $\boldsymbol{\varphi}$ to minimize $\mathcal{L}_\varphi$.

Vanilla SSL directly and only uses the predicted label of the network as $\hat{\boldsymbol{y}}_\varphi$ to self-tune the network. This setting has shown limited success [31]. The main drawback is that the tuned network is unable to correct its own mistakes and it may amplify the error. Since $\hat{\boldsymbol{y}}_\varphi$ is a pseudo label of $\boldsymbol{\varphi}$ and often over- or under-estimates the ground truth, how to generate the pseudo label properly and define an associated loss $\mathcal{L}_\varphi$ are important for SSL. Plausible designs of the label proxy $\hat{\boldsymbol{y}}_\varphi$ and the self-supervised loss $\mathcal{L}_\varphi$ should guide the SSL to move towards a latent convergence and therefore improve the segmentation. In this work, as explained in section II-C, we propose to encode adaptive shape knowledge as $\hat{\boldsymbol{y}}_\varphi$ to effectively drive the self-learning procedure.

C. EVOLVING SHAPE PRIOR

Vision system of human beings can perceive appearance-invariant structural information of the same object class across different imaging conditions. Shape prior shared by objects plays an important role in guiding the vision system to conduct case-specific customization. Statistical shape models are robust and often provide clinically more plausible segmentation than classification methods [32], [33]. Shape prior is even compelling for cross-modality segmentation where different modalities have very different imaging preferences [25]. These observations motivate us to investigate the feasibilities of exploring shape prior as a supervision signal in the SSL.

There are attempts to embed shape prior into deep networks by taking statistical shape model as constraints on network predictions [34], [35] or as regularizations on feature maps [36], [37]. However, effectively incorporating statistical shape model with our SSL scheme is not trivial. First, the fitting process of statistical shape models is often driven by handcrafted features which are not robust against the image appearance shift and the shape deficiency in network predictions. Second, shape prior should be adaptive and be able to synchronize its change with the iterative SSL procedure, thus providing accurate supervision to guide the learning procedure. Nevertheless, the fitting process of statistical shape model and the training of deep network are often optimized with different goals [35], how to synchronize them is still an open problem.

In this work, we propose a novel Evolving Shape Prior (ESP) to circumvent the problems. As shown in Fig. 3, the ESP serves as the $\hat{\boldsymbol{y}}_\varphi$ in SSL to provide adaptive shape prior as supervision. It is seamlessly implanted in the scheme and independent of feature extraction. Network predictions stimulate the recursive evolution of ESP and thus synchronize it with our SSL based case adaptation. Different from the population-restricted shape models, ESP provides case-specific shape knowledge for case adaptation. It has two key components: (1) a good starting point to trigger the SSL, (2) an evolving strategy in order to lead the SSL to stable model convergence.

1) MEAN SHAPE PRIOR FOR INITIALIZATION

Good initialization is crucial in triggering the SSL. We propose to initialize the ESP with a mean shape prior (MSP) map, denoted as $\boldsymbol{P}_{msp}$. Value at the location $\boldsymbol{v}$ (2D or 3D) in MSP represents the probability of $\boldsymbol{v}$ belonging to a pre-defined class C. Similar to [36], we estimate the MSP by computing the pixel- or voxel-wise proportion of each class C based on the ground truth labels $\boldsymbol{y}_s$ in source domain $\mathcal{S}$. The formulation of $\boldsymbol{P}_{msp}$ is

$$\boldsymbol{P}_{msp}(C|\boldsymbol{v}) = \frac{1}{S}\sum_{s=1}^{S} \mathbb{1}_C(\boldsymbol{y}_{s,\boldsymbol{v}}) \quad (4)$$

where S is the number of training images in source domain, $\mathbb{1}_C(\boldsymbol{y}_{s,\boldsymbol{v}})$ is an indicator function which returns 1 when $\boldsymbol{y}_{s,\boldsymbol{v}} = C$ and 0 otherwise. In this work, we verify our framework with binary segmentation tasks, i.e. foreground and background, and thus $C \in \{0, 1\}$. MSP foreground maps of the 8 datasets involved in this work (Fig. 2) are illustrated in Fig. 5. MSP maps present high similarity in

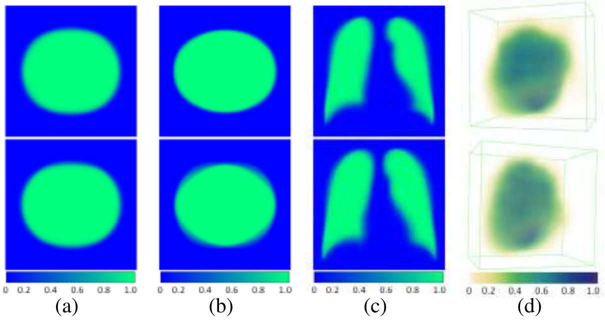


FIGURE 5. Mean shape prior foreground maps of 8 datasets (see Fig. 2 for details): (a) fetal abdomen in ultrasound, (b) fetal head in ultrasound, (c) human lung in X-ray, (d) whole heart in CT (top row) and MR (bottom row).

shape across datasets. All the images and labels are roughly cropped around their centers of mass to ensure an approximate alignment in location. MSP maps capture the main shape mode of each task but discard minor deformations (Fig. 5 (a)(b)(c)). Since the variations of heart structure and angle in 3D are very complex, MSP maps of heart in CT and MR are rough but still show proper hints of heart spatial configuration (Fig. 5(d)). As we prove in section III, MSP, as $\hat{\boldsymbol{y}}_\varphi$ for unseen images, effectively guides the case adaptation to produce initial segmentation with plausible shape completeness and hence builds a solid foundation for the followed ESP evolution. Case-specific shape prior will then be generated during evolution.

2) EVOLVING SHAPE PRIOR TOWARDS CONVERGENCE

Since our proposed SSL is only supervised by the ESP, making the ESP case-specific and accurateLY becomes crucial. In this work, we devise a novel evolution strategy to successively modify the ESP and fit it with the learning process. We hypothesize that, initialized with the $\boldsymbol{P}_{msp}$, the ESP will supervise the fine-tuning process to generate gradually refined segmentation. Segmentation in each iteration should contain some case-specific information and thus should be explored to amend the ESP. The ESP map, denoted as $\boldsymbol{P}_{esp}$, can then be enhanced and in turn provides better supervision to drive the network to produce more precise segmentation. This loop will be terminated after convergence. Key step in this loop is to effectively amend $\boldsymbol{P}_{esp}$ with the network prediction in each iteration, and further ensure a positive feedback in the loop to improve the segmentation. To achieve this, as shown in Fig. 3, we propose the following recursive evolution strategy,

$$\mathcal{D}_t = \mathop{\arg\max}_{\mathcal{D}'\subset\mathcal{S},|\mathcal{D}'|=K} \sum\nolimits_{\boldsymbol{d}\in\mathcal{D}'} DSC(\boldsymbol{q}^t,\boldsymbol{d}) \tag{5}$$

$$\boldsymbol{P}^t_{ame} = \sum_{k=1}^{K} \theta_k \boldsymbol{y}_k,\ \boldsymbol{y}_k \in \mathcal{D}_t, \quad s.t. \sum_{k=1}^{K} \theta_k = 1 \tag{6}$$

$$\boldsymbol{P}^t_{esp} = \alpha \boldsymbol{P}^{t-1}_{esp} + \beta \boldsymbol{P}^t_{ame}, \quad t \in \mathbb{Z}^+ \tag{7}$$

where t is iteration index, $\boldsymbol{q}^t$ is the segmentation in iteration t. $\boldsymbol{q}^t$ serves as a *query* image and is input into a comparator to retrieve the top K similar labels in $\mathcal{S}$ to form a *support set* $\mathcal{D}_t$. K is set as 3 to avoid outlier influence. We compare the similarity between $\boldsymbol{q}^t$ and label candidate $\boldsymbol{d}$ with Dice Similarity Coefficient (DSC). Label $\boldsymbol{y}_k$ in $\mathcal{D}_t$ are weighted with θ_k according to their DSC values and merged into an image $\boldsymbol{P}^t_{ame}$. $\boldsymbol{P}^{t-1}_{esp}$ then evolves into $\boldsymbol{P}^t_{esp}$ with the weighted amendment from $\boldsymbol{P}^t_{ame}$. $\boldsymbol{P}^0_{esp}$ is initialized with $\boldsymbol{P}_{msp}$.

The proposed strategy is simple but effective with three advantages. First, without any extra manual annotations, amending ESP with the evolving *support set* $\mathcal{D}_t$ can simulate an image-specific, visual-plausible shape supervision to strongly guide the fine-tuning. Second, DSC metric is utilized to find the most similar labels $\boldsymbol{d}$ in $\mathcal{S}$ to the intermediate segmentation to prevent the strict image alignment, which may be unstable or require additional annotation for transform matrix calculation. Third, with this evolving strategy, the network segmentation and the ESP will simultaneously move towards a similar shape. The SSL can thus stably achieve model convergence in only a few iterations.

For the self-supervised loss $\mathcal{L}_\varphi$ in SSL (Eq. 3), we customize it as a regression loss, i.e., $\mathcal{L}_\varphi = \| \hat{\boldsymbol{p}}^t - \boldsymbol{P}^t_{esp} \|_1$. $\hat{\boldsymbol{p}}^t$ is the probability map predicted by network in iteration t. The main reason behind this design is that, as a fusion result of retrieved similar shape labels, $\boldsymbol{P}^t_{esp}$ is often blurry and its soft version tends to have more shape clues than its binary label version. Therefore, regression based loss is better than binary label based cross-entropy in exploring $\boldsymbol{P}^t_{esp}$.

3) FURTHER ENHANCE THE STABILITY

In our practice, we find that ESP based SSL tends to suffer from sudden performance drops during iteration in some challenging tasks, like the whole heart segmentation. We interpret this phenomenon as that, although the evolving shape prior $\boldsymbol{P}^t_{esp}$ and the segmentation $\boldsymbol{q}^t$ can be very close, there still exist some discrepancies between them in tough tasks. The segmentation may sometimes be closer to the latent ground truth than $\boldsymbol{P}^t_{esp}$. In this case, forcing the network to fit the $\boldsymbol{P}^t_{esp}$ is destructive. Therefore, we propose to further enhance the stability of SSL by directly introducing the segmentation $\boldsymbol{q}^t$ into the basic ESP (Eq. 7) as a balance,

$$\boldsymbol{P}^t_{esp} = \alpha \boldsymbol{P}^{t-1}_{esp} + \beta \boldsymbol{P}^t_{ame} + \gamma \boldsymbol{q}^t \tag{8}$$

III. EXPERIMENTAL RESULTS

A. IMPLEMENTATION DETAILS

We implemented our work in *Tensorflow*, using an NVIDIA GeForce GTX TITAN Xp GPU (12GB). In training networks on source corpus, we update the weights of networks with an Adam optimizer (batch size=2, initial learning rate is 0.001, momentum term is 0.5, total iteration=8000). During the case adaptation, we update the weights of all networks with smaller initial learning rates as 0.0001. In adaptation, We update the weights of networks 3 times referred to $\mathcal{L}_\varphi$ in each iteration to properly fit current ESP. θ_k is set to *0.5*, *0.3* and *0.2* for top 3 candidates in Eq. 6, respectively. For each image, our method is fast and only needs less than 25 iterations (*about 0.2 sec/iteration for 2D task, 2 sec/iteration for 3D task*) before achieving a satisfying segmentation and stable convergence, less than 10 seconds in 2D task and 1 minute in 3D task of the whole initialization and evolution process. Necessary data augmentation, including flipping and rotation, are conducted as default. All tasks in this paper share the same parameter setting which proves to be acceptable across all verifications on source corpora.

B. EVALUATION CRITERIA AND METHOD COMPARISON

We adopt 5 metrics to evaluate the segmentation results from both shape similarity and boundary similarity perspectives, including DSC (%), Conformity (%) [38], Jaccard (%), Hausdorff Distance of Boundaries (HDB, [pixel] or [voxel]) and Average Distance of Boundaries (ADB, [pixel] or [voxel]). For verification, we conduct extensive comparisons among different methods and ablation studies. For clarity of description, in this paper, we quote all methods with a '*A2B*' naming format, in which '*A*' is the corpus where the source model is trained, '*B*' is the corpus in which each image is directly tested or the subject that case adaptation is applied to. *A2A* thus means, the model is trained with training dataset of *A* and tested with the testing dataset of *A*. Histogram matching (*HistM*) and CycleGAN [39](*CyGAN*) are considered in this study for comparison. *HistM* adjusts the histogram of a floating image to a reference one. It is a classic method in dealing with intensity distribution variations [40]. In this paper, we define the reference histogram (128 bin) as the averaged histogram over a training dataset. *A2B-HistM* means that every testing image in *B* is firstly aligned to the reference histogram of *A* and then tested by the segmentor of *A*. CycleGAN produces realistic appearance translation between different image domains [9]. It has high potentials in unifying image appearances. *A2B-CyGAN* means that, each testing image in *B* is firstly translated into an *A*-like appearance by a pre-trained mapping between *A* and *B*, and then tested by the segmentor of *A*. Training of CycleGAN is time-consuming and needs two balanced corpora from *A* and *B*.

For ablation study, we mainly compare four variations of case adaptation: (1) *A2B-vSSL*, vanilla self-supervised learning (*vSSL*), which is depicted in section II-B. *A2B-vSSL* means that, segmentor of *A* firstly deploys on the testing image from *B* and then is directly and solely fine-tuned by the generated binary label. (2) *A2B-MSP*, standing for the fine-tuning only supervised by the initial, static MSP. With more complete shape supervision, *A2B-MSP* may perform better than *A2B-vSSL* which often suffers from its own mistakes. (3) *A2B-ESP*, which is our proposed method. It should output much better results than *A2B-MSP*. All *A2B-ESP* methods adopt $\alpha = 0.6$ and $\beta = 0.4$ in Eq. 7. (4) *A2B-EHCE*, which is the stability-enhanced version of *A2B-ESP* (Eq. 8). All *A2B-EHCE* methods adopt $\alpha = 0.3$, $\beta = 0.2$ and $\gamma = 0.5$ in Eq. 8. All case adaptation methods iterate on each image for 25 iterations. Only the results in iteration 25 are reported. We keep reasonable settings for all methods for fair comparisons.

TABLE 1. Method Comparison of Fetal Abdomen Segmentation in Ultrasound Image (Direction: *A2B*).

Method	Metrics				
	DSC	Conformity	Jaccard	Hdb	Adb
C2C	94.894	89.097	90.382	46.991	9.3761
C2S	90.216	77.866	82.397	58.288	14.303
C2C-MSP	95.892	91.358	92.157	14.523	4.4780
C2C-ESP	97.214	94.225	94.606	11.930	3.0535
C2S-HistM	85.137	62.222	74.847	67.814	18.722
C2S-CyGAN	91.623	81.374	84.729	67.548	15.328
C2S-vSSL	91.202	80.359	84.020	43.985	11.525
C2S-MSP	96.134	91.884	92.604	17.651	4.4312
C2S-ESP	97.037	93.848	94.274	15.449	3.4569

TABLE 2. Method Comparison of Fetal Abdomen Segmentation in Ultrasound Image (Direction: *B2A*).

Method	Metrics				
	DSC	Conformity	Jaccard	Hdb	Adb
S2S	91.816	81.765	85.088	70.633	16.374
S2C	89.877	76.687	81.984	75.028	19.118
S2S-MSP	96.075	91.749	92.501	13.024	4.2698
S2S-ESP	96.982	93.743	94.162	11.908	3.2939
S2C-HistM	87.675	70.901	78.446	77.170	21.310
S2C-CyGAN	89.328	75.507	80.997	71.109	18.192
S2C-vSSL	93.033	84.821	87.094	43.322	10.073
S2C-MSP	95.742	91.035	91.877	14.655	4.6201
S2C-ESP	96.654	93.026	93.558	14.229	3.6960

C. QUANTITATIVE AND QUALITATIVE ANALYSIS

We verify our proposed framework with four distinctive tasks where segmentation performance gaps occur due to different factors. Each task contains two datasets with recognizable appearance shift (Fig. 2). We conduct extensive experiments with a strict bi-directional manner, i.e., *A2B* and *B2A*, for thorough tests.

1) ULTRASOUND IMAGE OF FETAL ABDOMEN

We collected ultrasound images of fetal abdomen to verify the gap using different scanners. Abdomen circumference is important for fetal weight estimation. In total, 1540 images were acquired using a Sonoscope C1-6 ultrasound scanner (denoted as *C*) with gestational age from 30 to 34 week. 1515 images were acquired using a Siemens Acuson Sequoia 512 ultrasound scanner (denoted as *S*), with gestational age from 24 to 40 week. Free deformation and fetal pose were allowed during image acquisition (Fig. 2(a)). In both datasets, we randomly select 900 images with augmentation for training, the rest for testing. Experienced experts provide the segmentation ground truth. All images are cropped to center around the fetal abdomen region and resized to the size of 320 × 320.

In Table 1 and 2, models trained and tested in source corpora (*C2C* and *S2S*) achieve acceptable results [41], but are degraded when applied to new corpora (*C2S* and *S2C*). Histogram matching and CycleGAN only bring about slight improvement (*C2S-HistM*/*C2S-CyGAN* vs *C2S*) and even make it worse (*C2S-HistM* vs *C2S*, *S2C-HistM*/*S2C-CyGAN* vs *S2C*). This may indicate that corpus-based intensity distribution and appearance style are not specific enough in case adaptation. Case adaptation based variations generally

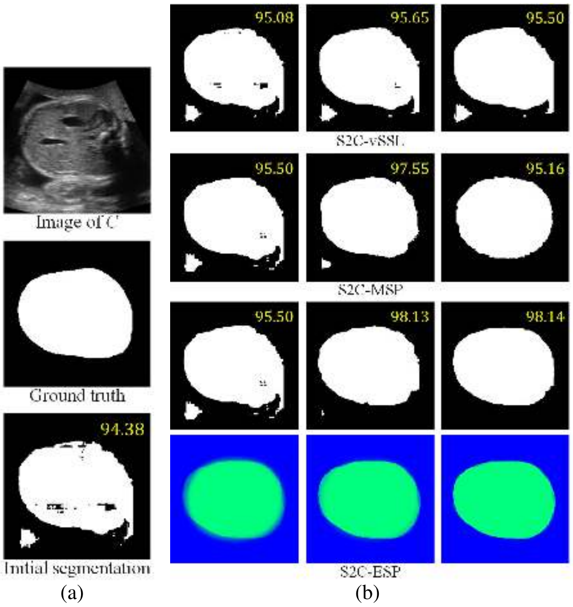


FIGURE 6. (a) From top to bottom, an ultrasound image of fetal abdomen from corpus *C*, segmentation ground truth and the initial segmentation result produced by the segmentor trained on corpus *S*, (b) From top to bottom, intermediate segmentation results of *S2C-vSSL*, *S2C-MSP*, *S2C-ESP* and *ESP* images at iteration 1, 8 and 25, respectively. Yellow digits denote the DSC measurement of each segmentation result. Color version for better view.

present better results than competitors as the iteration propagates. Among them, *vSSL* already get notable improvement (*S2C-vSSL* vs *S2C*). However, since vanilla SSL based methods (*C2S-vSSL*, *S2C-vSSL*) are supervised by its own prediction, they often tend to stick on some wrong candidates, as shown in Fig. 6 (b). *MSP* is able to imporve the sgementation (about 6% DSC in both *C2S-MSP* vs *C2S* and *S2C-MSP* vs *S2C*) as it encodes complete shape information. Although MSP helps in eliminating false positives, it sometimes makes segmentation converge to the shape prior (Fig. 6(b)). *ESP* based methods produce best results in the bidirectional tests (about 0.9% in DSC over *MSP* based methods), in all evaluated metrics. From Fig. 6 (b), we can see that, *S2C-ESP* gradually generates the refined segmentation as iteration increases. At the same time, the *ESP* image also evolves to a case-specific pattern from its prototype (Fig. 5 (a)) to closely match the learning procedure. We also notice improvements in *C2C-MSP*/*C2C-ESP* vs *C2C* and *S2S-MSP*/*S2S-ESP* vs *S2S*. We owe this to the capacities of *MSP*/*ESP* in guiding networks to suppress false positives and overcome intra-corpus appearance variations which may be caused by different imaging parameters.

2) ULTRASOUND IMAGE OF FETAL HEAD

We also collected ultrasound images of the fetal head. Its circumference is another basic measurement in prenatal examination. Besides different scanners, datasets here blend a factor of different timepoints where fetal head shape does not change much, but the appearance of its inner structures, like the thalamus and cerebellum, changes dramatically due to the fetal growth and increasing acoustic shadows (Fig. 2(b)). 1315 images were acquired using a Sonoscope C1-6 ultrasound scanner (denoted as *C*) with gestational age from 18 to 24 week. 1372 images were acquired using a Siemens Acuson Sequoia 512 ultrasound scanner (denoted as *S*), with gestational age from 24 to 40 week. The training and testing are set like that in section III-C1.

TABLE 3. Method Comparison of Fetal Head Segmentation in Ultrasound Image (Direction: *A2B*).

Method	Metrics				
	DSC	Conformity	Jaccard	Hdb	Adb
C2C	94.254	86.865	89.623	63.527	15.378
C2S	88.583	72.528	80.091	69.218	19.270
C2C-MSP	97.132	94.036	94.464	12.070	3.1843
C2C-ESP	98.030	95.951	96.153	10.081	2.2050
C2S-HistM	88.144	71.516	79.364	71.735	20.437
C2S-CyGAN	94.160	87.219	89.188	71.325	17.048
C2S-vSSL	92.663	83.926	86.470	51.519	11.873
C2S-MSP	96.410	92.485	93.113	15.013	4.0131
C2S-ESP	97.630	95.126	95.378	13.052	2.6752

TABLE 4. Method Comparison of Fetal Head Segmentation in Ultrasound Image (Direction: *B2A*).

Method	Metrics				
	DSC	Conformity	Jaccard	Hdb	Adb
S2S	95.908	91.313	92.241	61.312	12.090
S2C	69.221	1.4783	54.229	91.950	29.708
S2S-MSP	96.220	92.073	92.761	11.884	4.1086
S2S-ESP	97.944	95.782	95.981	7.1835	2.1989
S2C-HistM	70.927	8.9883	56.270	92.028	29.297
S2C-CyGAN	91.968	81.870	85.480	75.449	20.310
S2C-vSSL	86.329	67.018	76.405	83.717	25.094
S2C-MSP	97.283	94.366	94.738	20.261	3.3152
S2C-ESP	97.634	95.121	95.396	19.960	2.9840

Suffering from the complicated appearance shift caused by the composite effect of scanners and timepoints, as observed in Table 3 and 4, much larger gap than that in section III-C1 emerges when we compare *C2C* vs *C2S* (6% DSC drop) or *S2S* vs *S2C* (30% DSC drop). Imbalance between the bidirectional drops also shows that segmentors trained on each corpus captures different features, which then results in the difference of segmentor robustness. Histogram matching here has almost no contribution to the segmentation (*C2S-HistM* vs *C2S*, *S2C-HistM* vs *S2C*), while CycleGAN achieves about 6% and 22% DSC improvement on the two directions, respectively. This indicates that, when different scanners and timepoints are blended in imaging, the appearance shift is beyond what the HistM can capture, but can be partially addressed by CycleGAN. Case adaptation based methods continue to present improvements and better results than its competitors, except the vanilla *SSL* (*C2S-vSSL* vs *C2S-CyGAN*, *S2C-vSSL* vs *S2C-CyGAN*). *ESP* based methods significantly close the gap and achieve comparable performance to the model trained on the source corpora (*C2S-ESP* vs *C2C-ESP*, *S2C-ESP* vs *S2S-ESP*).

In Fig. 7, we visualize the DSC improvement of *C2S-ESP* over initial segmentation along with iteration on 472 images from *S* (Fig. 7 (a)), and *S2C-ESP* on 415 images from *C* (Fig. 7 (b)). We can see that, *ESP* based methods brings about improvement for almost all the testing cases. The maximum

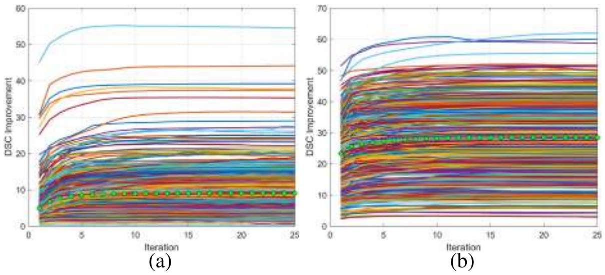


FIGURE 7. Curves of DSC improvement over initial segmentation. (a) *C2S-ESP* on 472 images from *S*, (b) *S2C-ESP* on 415 images from *C*. Green dot denotes the averaged DSC improvement at each iteration.

TABLE 5. Method Comparison of Lung Segmentation in X-ray Image (Direction: *A2B*).

Method	Metrics				
	DSC	Conformity	Jaccard	Hdb	Adb
J2J	97.497	94.855	95.124	13.639	1.3673
J2M	63.525	-24.492	0.4734	48.375	10.232
J2J-MSP	96.478	92.636	93.235	16.418	1.7069
J2J-ESP	97.069	93.932	94.322	17.165	1.5527
J2M-HistM	80.429	47.046	68.512	42.907	7.8955
J2M-CyGAN	91.916	81.959	85.280	23.391	3.9601
J2M-vSSL	74.851	23.578	61.575	44.704	8.9772
J2M-MSP	90.614	79.140	82.912	25.177	4.6339
J2M-ESP	93.040	84.872	87.082	21.095	3.3910

DSC improvement is about 55% in *C2S-ESP*, and 60% in *S2C-ESP*. The curves of *S2C-ESP* spans a larger range than that of *C2S-ESP*, which may indicate that appearance variation in *C* is much larger than *S*. With the evolution, almost all curves plateau after only about 5 iterations (*about 1 second*) and then keep the trend.

3) X-ray IMAGE OF CHEST

We investigate the segmentation performance gap on X-ray images of chest acquired in different sites with different protocols. Lung in the chest has much more complex shape than fetal abdomen and head. Two publicly available datasets are included [42]. *JSRT Set*: a set compiled by the Japanese Society of Radiological Technology (denoted as *J*). It contains 247 images, 154 have lung nodules (100 malignant cases, 54 benign cases), and 93 have no nodules. We randomly split the dataset into 200/47 for training/testing. *Montgomery Set*: a set from the Department of Health and Human Services, Montgomery County, Maryland (denoted as *M*). It consists of 138 images, 80 of them are normal and 58 are abnormal with manifestations of tuberculosis. We randomly split the dataset into 90/48 for training/testing. Both training datasets are augmented to 600 with proper rotation and deformation. We treat left and right lungs as the same class. All images are cropped to center around the lung and resized to 256 × 256.

As shown in Fig. 2 (c), these two sets have different appearances from each other. In Table 5 and 6, large segmentation performance drops occur in both directions (34% DSC in *J2J* vs *J2M*, 24% DSC in *M2M* vs *M2J*). Histogram matching contributes to the best results with *M2J-HistM* in Table 6. This indicates that the main appearance difference between these two sets may be intensity distribution variations. However, since histogram matching is very sensitive to the reference histogram which is prone to be affected by the global content of reference images, it therefore only brings plain improvement in *J2M-HistM*, see Fig. 8. With a stable and realistic style translation between two sets, CycleGAN based methods improve the segmentation by more than 20% DSC in both directions (*J2M-CyGAN* vs *J2M*, *M2J-CyGAN* vs *M2J*). However, as instanced in Fig. 8, improvement provided by CycleGAN is limited, since the translation results are subject to content blurry and boundary shift [12]. Compared to competitors, vanilla *SSL* methods obviously show no advantages (*J2M-vSSL*, *M2J-vSSL*). Same as Fig. 6, vanilla *SSL* can improve segmentation but tends to be trapped by suboptimal segmentation, and the risk is intensified by the complex shape of lung. Although there still exists a gap of about 4% in DSC, *ESP* introduces almost the best results among all methods and greatly narrows down the initial gap (*J2M-ESP*, *M2J-ESP*). Segmentation is nearly maintained when apply *ESP* to *J2J*/*M2M*, this might be the guidance from *ESP* is subtle when original segmentation achieves good performance (DSC is *97.49* in *J2J*, *96.76* in *M2M*).

TABLE 6. Method Comparison of Lung Segmentation in X-ray Image (Direction: *B2A*).

Method	Metrics				
	DSC	Conformity	Jaccard	Hdb	Adb
M2M	96.762	93.165	93.835	21.821	2.1210
M2J	72.343	11.434	58.530	39.917	9.4631
M2M-MSP	96.381	92.415	93.066	20.453	2.1810
M2M-ESP	96.928	93.580	94.094	19.070	1.8500
M2J-HistM	95.091	89.620	90.680	26.468	2.9847
M2J-CyGAN	92.665	84.101	86.374	31.689	4.4092
M2J-vSSL	84.883	63.506	74.075	51.784	10.855
M2J-MSP	93.141	85.161	87.227	30.547	3.8320
M2J-ESP	93.465	85.887	87.809	25.180	3.3617

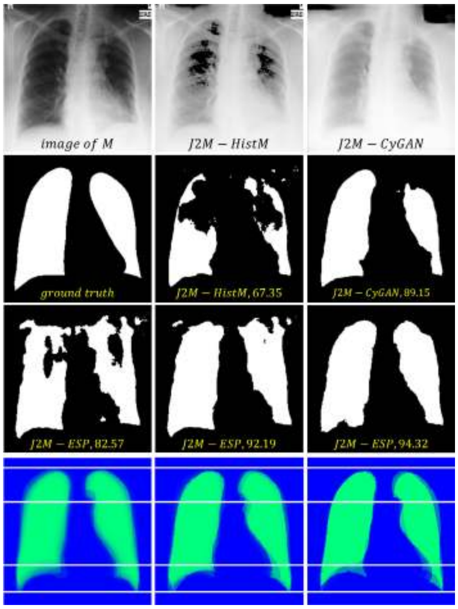


FIGURE 8. Visual comparison among histogram matching, CycleGAN and *ESP*. Results of *ESP* at iteration 1, 3 and 25 are shown. Yellow digits denote DSC value, white lines on *ESP* images for reference. Best view in color version.

TABLE 7. Method Comparison of MYO, LV and RV Segmentation in cine-MR images (Direction: *A2B*).

Method	MYO		LV		RV	
	DSC	Hdb	DSC	Hdb	DSC	Hdb
M2M	91.785	3.452	86.184	4.680	88.336	6.433
M2A	90.498	4.732	83.718	6.522	85.546	7.080
M2M-MSP	92.602	3.148	87.026	4.553	89.417	5.591
M2M-ESP	92.894	3.101	87.256	4.528	90.152	5.405
M2A-HistM	89.805	4.865	82.698	6.489	85.507	7.324
M2A-CyGAN	90.493	4.873	82.877	6.616	85.448	7.154
M2A-vSSL	90.748	4.641	83.711	6.445	85.737	7.184
M2A-MSP	91.451	4.411	84.405	6.335	86.503	7.365
M2A-ESP	91.754	4.277	84.581	6.131	87.255	7.489

In Fig. 8, we visually compare the histogram matching, CycleGAN and *ESP* in improving the segmentation of an image from set *M*. Segmentor trained on *J* only achieves an initial DSC of *60.662%*. Translation of *J2M-HistM* and *J2M-CyGAN* improve the segmentation but have obvious flaws. *J2M-ESP* starts with a poor prediction (iteration 1, DSC *82.57%*), but significantly refines the segmentation as learning iterates (DSC *92.19%* at iteration 3, *94.32%* at iteration 25). False positives and negatives are gradually rectified and finally a clean and complete shape is presented. The associated *ESP* image also evolves accordingly to guide the refinement, although the ESP is not exactly matched with the ground truth. Reference lines are plotted for readers to compare the evolution details.

4) cine-MR IMAGES OF HEART

We investigate the effectiveness of our method in bridging multi-category segmentation performance gap on cine-MR images of heart acquired from different clinical centres and vendors. We perform the comparison experiments on two public datasets including *M&Ms* [43] and *ACDC* [44]–[46]. The *M&Ms* dataset (denoted as *M*) was collected in three different countries using three different magnetic resonance scanner vendors.[1] It contains 3284 slices from 150 patients and we randomly split the dataset into 2332/952 for training/testing according to patient. *ACDC* dataset (denoted as *A*) was from the University Hospital of Dijon. It contains 1902 slices from 200 patients. We randomly split the dataset into 1342/560 for training/testing according to patient. Both of the datasets contain three target segmentation categories including left (LV) and right ventricle (RV) blood pools, as well as the left ventricular myocardium (MYO). The training datasets are augmented with flipping and 90-degree rotation. All the slices are cropped to center around the heart and resized to 128 × 128.

TABLE 8. Method Comparison of MYO, LV and RV Segmentation in cine-MR images (Direction: *B2A*).

Method	MYO		LV		RV	
	DSC	Hdb	DSC	Hdb	DSC	Hdb
A2A	93.921	3.670	88.907	5.011	90.080	6.589
A2M	86.258	6.428	81.061	7.029	84.565	8.497
A2A-MSP	94.911	2.756	89.617	4.737	90.544	6.575
A2A-ESP	94.970	2.758	89.710	4.402	90.740	6.487
A2M-HistM	85.144	6.070	79.127	7.135	81.112	8.352
A2M-CyGAN	86.045	6.329	80.853	7.296	84.197	8.441
A2M-vSSL	86.935	5.385	81.744	6.251	85.627	7.114
A2M-MSP	87.785	5.137	82.317	6.158	86.522	6.477
A2M-ESP	87.844	5.012	82.591	6.021	87.136	6.405

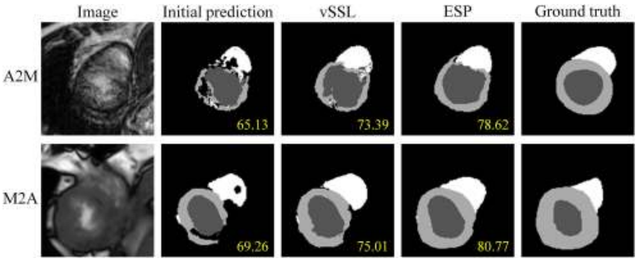


FIGURE 9. Visual comparison among initial prediction, vSSL and ESP. The numbers on the bottom right represent DSC of each method.

As shown in Fig. 2 (d), there are large appearance differences between the two datasets. From Table 8 and Table 7, we can observe that the segmentation performance suffers drop in most of the tissues in both directions. The A2M direction occurs larger decline than M2A direction, e.g. 1% DSC drop of MYO in M2A vs. 8% in A2M. This is because a more generalized model can be trained in the dataset *M* acquired from multiple data sources. *HistM* and *CyGAN* fail in boosting the performance of the baseline due to the complex appearance differences. Particularly, the sensitivity of *HistM* makes it unable to perform stable appearance transfer, resulting in worse results. Owing to the effectiveness of the SSL, the *vSSL* can bring slight improvement for the baseline in most of the tissues. After embedding the strong shape prior by *MSP* and *ESP*, the shape prior based SSL methods can improve the baseline in all of the three tissues of both directions. As shown in Fig. 9, the *ESP* contributes the best segmentation performance compared with the *vSSL* and initial prediction. We can also observe that the *ESP* method can repair the tattered segmentation map of initial predictions by introducing the shape prior based supervision. It can be proved that our proposed methods can not only bridge the segmentation performance gap in single category but also boost the accuracy of the multi-category segmentation.

5) CT AND MR VOLUME OF HEART

We finally address the most challenging task of cross-modality whole heart segmentation. Multi-Modality Whole Heart Segmentation Challenge 2017 datasets are used in this study [47]. Datasets consisting of 60 CT and 60 MR volumes (20/40 for training/testing) are acquired from different patients without alignment. Seven heart substructures are labeled by experts. As testing labels are held by Challenge organizers and unavailable, we split training datasets of CT (denoted as *T*) and MR (denoted as *R*) into 12/8 for our training/testing. Training and testing parts of *T* and *R* are augmented with rotation and deformation to 228 and 72, respectively. In this work, we treat the 7 substructures as the same class. All volumes are cropped to roughly center around the whole heart and resized to 96 × 96×64. The whole volume serves as our network input, rather than a patch manner.

[1] We only used the released annotated training set of this public dataset.

TABLE 9. Method Comparison of Whole Heart Segmentation in CT/MR Volume (Direction: *A2B*).

Method	Metrics				
	DSC	Conformity	Jaccard	Hdb	Adb
T2T	90.706	79.197	83.153	18.841	1.3112
T2R	51.998	-109.23	36.009	20.710	3.1357
T2T-ESP	91.427	81.022	84.330	19.305	1.0237
T2R-CyGAN	76.208	36.980	61.692	23.320	3.3242
T2R-vSSL	62.283	-33.546	46.222	33.875	7.7229
T2R-MSP	70.225	11.064	54.556	21.440	3.3007
T2R-ESP	82.142	55.620	69.967	17.445	2.0160
T2R-EHCE	83.346	59.490	71.630	17.173	1.7907

TABLE 10. Method Comparison of Whole Heart Segmentation in CT/MR Volume (Direction: *B2A*).

Method	Metrics				
	DSC	Conformity	Jaccard	Hdb	Adb
R2R	89.685	76.784	81.410	12.871	1.1647
R2T	76.704	38.161	62.447	22.644	3.8082
R2R-ESP	89.933	77.401	81.816	16.009	0.9590
R2T-CyGAN	56.111	-252.66	42.117	26.135	3.1161
R2T-vSSL	68.184	4.0334	52.056	28.780	6.1670
R2T-MSP	72.812	22.175	57.744	26.080	2.3811
R2T-ESP	79.722	48.470	66.457	24.863	2.1509
R2T-EHCE	80.930	52.478	68.087	23.057	2.4991

Fig. 2 (e) demonstrates the very different appearances of CT (top row) and MR (bottom row) in imaging the heart, especially the intensity and contrast of chambers and myocardia. In Table 9, the big difference is reflected by about 40% DSC drop from *T2T* to *T2R*. However, the DSC drop from *R2R* to *R2T* is relatively small, only 13% (Table 10). Better contrast between heart and surrounding tissues in CT than that in MR may lead to the imbalance. This imbalance is also observed in the results of CycleGAN. Implemented in 2D, synthesizing MR-like slices from CT (*R2T-CyGAN*) seems to be much harder than the contrary (*T2R-CyGAN*) [11]. Stuck in its own mistakes, especially in 3D applications, vanilla *SSL* performs worst among all case adaptation methods and becomes even worse than the baseline (*R2T-vSSL* vs *R2T*). *MSP* also tends to lose power in a 3D scenario where complex geometry, rotation and deformation are involved. It only facilitates the segmentation with *T2R-MSP*, while degrades the performance in *R2T-MSP*. Improvements in both directions firstly occur with *ESP*, about 30% DSC increment from *T2R* to *T2R-ESP* and 3% DSC from *R2T* to *R2T-ESP*. However, as illustrated in Fig. 10, the performance of self-supervised learning process guided by *ESP* may suddenly decline on some difficult volumes. The discrepancy between the accuracy-increasing prediction of network and the estimated *ESP* map is the main course. Network prediction should thus be properly incorporated into *ESP* to calibrate the supervision signal. As defined in Eq. 8, the enhanced *ESP* finally achieves another 1% DSC improvement (*T2R-EHCE*, *R2T-EHCE*) by stabilizing the recursive case adaptation (Fig. 10).

DSC improvement curves of all fine-tuning methods on 72 testing volumes are shown in Fig. 10. *vSSL*, *MSP*, *ESP* and *EHCE* gradually present refined results and finally a fast, stable and positive convergence trend on almost all testing cases. At iteration 25, the maximum DSC improvement of *T2R-EHCE* is about 60%, while it is 12% of *R2T-EHCE*. Advantages of methods and differences between two testing directions are well visualized through the figures.

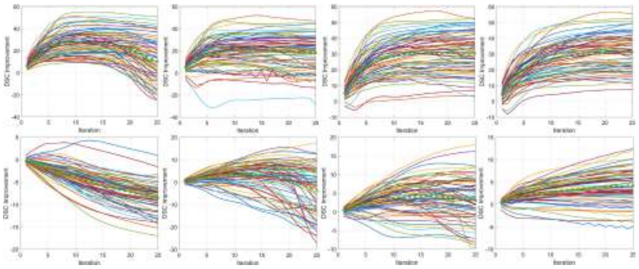


FIGURE 10. Curves of DSC improvement over initial segmentation. Top row: *T2R-vSSL*, *T2R-MSP*, *T2R-ESP*, *T2R-EHCE*. Bottom row: *R2T-vSSL*, *R2T-MSP*, *R2T-ESP*, *R2T-EHCE*. Green dot is averaged improvement at the iteration.

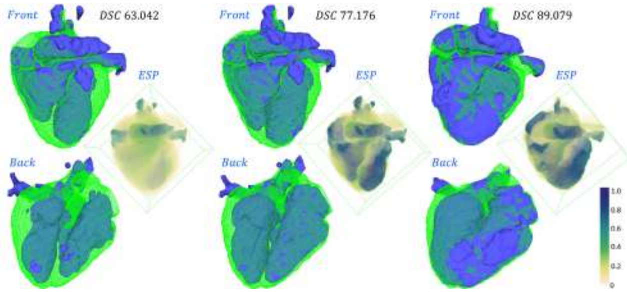


FIGURE 11. From left to right, intermediate segmentation and *ESP* output of *T2R-EHCE* at iteration 1, 5 and 25. Green mesh denotes ground truth, blue surface denotes segmentation. Color bar for *ESP*. Best view in color version.

In Fig. 11, we further visualize the intermediate results of *T2R-EHCE* in segmenting an MR volume from *M* (iteration 1, 5 and 25). Segmentor trained on CT volumes only achieve an initial DSC of *53.183%*. From both the front and back views, segmentation is significantly refined as iteration increases. High agreement is achieved between segmentation surface and ground truth mesh at iteration 25. The enhanced *ESP* volume also rapidly evolves to be sharp and segmentation-like from its initial fuzzy and rough form (Fig. 5 (d)). Our final segmentation results are not that perfect but are comparable with single-modality based methods [28]. These results proves the feasibility of direct cross-modality segmentation and provide good label initializations for further studies.

6) ANALYZING THE SELF-SUPERVISED LEARNING

To gain more insights about the SSL scheme in bridging segmentation performance gap, in Fig. 12, we visualize the maximum absolute change among all feature maps of each network layer along iteration. All layers in the network are countered. Specifically, we choose the *S2C-ESP* on a fetal head ultrasound image and *T2R-EHCE* on an MR volume as exemplars. As we can observe, significant and slight changes happen to layers in an interleaved way in both figures. This may indicate that, in case adaptation, some knowledge of

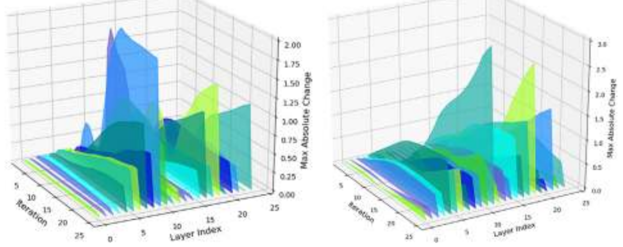


FIGURE 12. Layer-wise maximum feature map value change along iteration. Left: *S2C-ESP* on an ultrasound image, right: *T2R-EHCE* on an MR volume.

object is preserved across different imaging conditions while the rest is discarded and amended. Magnitude change of layers in *T2R-EHCE* is much larger than that in *S2C-ESP*. This verifies that cross-modality segmentation is more challenging and the dramatic changes in specific layers help networks bridge the large segmentation performance gap.

IV. LIMITATION AND DISCUSSION

We investigate the feasibility of self-supervised learning scheme in helping deep segmentation networks generalize to different imaging conditions. It is crucial before deep models can be embedded into real clinical workflow. Extensive experiments on eight datasets with typical imaging factors demonstrate the efficacy of our proposed framework. Without any burden on model complexity, our method narrows down the segmentation performance gap in an economical, stable and fast way. With detailed comparisons and visualizations, we hope to provide readers with insights into this emerging research direction.

Although being promising, there still exist several key points in our framework for future study. Currently, we are using DSC as a key metric to retrieve the most similar labels (Eq. 5) in source domain to amend ESP. However, DSC is still not accurate enough in describing the global and local similarity between segmentation and annotation label. Directly merging all retrieved labels based on the DSC values also ignores the importance of each pixel/voxel. Better label retrieval and fusion strategy should be considered [31]. Although the coefficients in *ESP* for evolution (Eq. 7, Eq. 8) are robust in almost all tasks, they are empirically set, which may limit the iterative refinement and cause the sudden drop as shown in Fig. 10. Learning to adaptively set these parameters is an interesting research direction. Furthermore, there is still a performance gap between our method and the state-of-art methods, e.g. about 4% DSC gap of our methods to the ranking top 1 approach in ACDC challenge [44]. Achieving a robust segmentation performance bridging is our target in the future. Finally, the diseased tissues, i.e. tumors, usually have irregular shapes, which may affect the performance of our methods. We will equip our method with the advanced technology, e.g. style transfer, to achieve a universal approach for bridging the segmentation gap. In these scenarios, generating effective *ESP* and properly coupling it with the adaptation procedure will be critical.

V. CONCLUSION

In this work, we attempt to narrow the segmentation performance gap encountered by deep networks under image appearance shift. The problem is very general as illustrated on eight typical datasets. We argue that *case adaptation* should be more tractable and be considered more than domain adaptation in solving the problem for real clinical scenarios. Our work integrates the strengths of traditional shape prior and self-supervised learning in a novel way. To the best of our knowledge, this is the first work exploring self-supervised learning for medical image segmentation. We extensively validate the proposed framework and provide diverse results with insights to prove that, *case adaptation* is lightweight, efficient and feasible in helping deep models bridge various segmentation performance gaps.

ACKNOWLEDGMENT

(Chaoyu Chen and Xin Yang contributed equally to this work.)

REFERENCES

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

[2] G. Hinton, ''Deep learning—A technology with the potential to transform health care,'' *J. Amer. Med. Assoc.*, vol. 320, no. 11, pp. 1101–1102, 2018.

[3] W. W. Stead, ''Clinical implications and challenges of artificial intelligence and deep learning,'' *J. Amer. Med. Assoc.*, vol. 320, no. 11, pp. 1107–1108, 2018.

[4] E. Gibson, Y. Hu, N. Ghavami, H. U. Ahmed, C. Moore, M. Emberton, H. J. Huisman, and D. C. Barratt, ''Inter-site variability in prostate segmentation accuracy using deep learning,'' in *Proc. MICCAI*. Cham, Switzerland: Springer, 2018, pp. 506–514.

[5] G. Csurka, ''Domain adaptation for visual applications: A comprehensive survey,'' 2017, *arXiv:1702.05374*. [Online]. Available: http://arxiv.org/abs/1702.05374

[6] X. Yang, H. Dou, R. Li, X. Wang, C. Bian, S. Li, D. Ni, and P.-A. Heng, ''Generalizing deep models for ultrasound image segmentation,'' in *Proc. MICCAI*. Cham, Switzerland: Springer, 2018, pp. 497–505.

[7] M. Drozdzal, G. Chartrand, E. Vorontsov, M. Shakeri, L. Di Jorio, A. Tang, A. Romero, Y. Bengio, C. Pal, and S. Kadoury, ''Learning normalized inputs for iterative estimation in medical image segmentation,'' *Med. Image Anal.*, vol. 44, pp. 1–13, Feb. 2018.

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, ''Generative adversarial nets,'' in *Proc. NIPS*, 2014, pp. 2672–2680.

[9] D. Nie, R. Trullo, J. Lian, L. Wang, C. Petitjean, S. Ruan, Q. Wang, and D. Shen, ''Medical image synthesis with deep convolutional adversarial networks,'' *IEEE Trans. Biomed. Eng.*, vol. 65, no. 12, pp. 2720–2730, Dec. 2018.

[10] Y. Huo, Z. Xu, S. Bao, A. Assad, R. G. Abramson, and B. A. Landman, ''Adversarial synthesis learning enables segmentation without target modality ground truth,'' in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 1217–1220.

[11] Z. Zhang, L. Yang, and Y. Zheng, ''Translating and segmenting multimodal medical volumes with cycle- and shape-consistency generative adversarial network,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9242–9251.

[12] C. Chen, Q. Dou, H. Chen, and P.-A. Heng, ''Semantic-aware generative adversarial nets for unsupervised domain adaptation in chest X-ray segmentation,'' 2018, *arXiv:1806.00600*. [Online]. Available: http://arxiv.org/abs/1806.00600

[13] K. Kamnitsas, C. Baumgartner, C. Ledig, V. Newcombe, J. Simpson, A. Kane, D. Menon, A. Nori, A. Criminisi, D. Rueckert, and B. Glocker, ''Unsupervised domain adaptation in brain lesion segmentation with adversarial networks,'' in *Proc. IPMI*. Cham, Switzerland: Springer, 2017, pp. 597–609.

[14] M. W. Lafarge, J. P. Pluim, K. A. Eppenhof, P. Moeskops, and M. Veta, "Domain-adversarial neural networks to address the appearance variability of histopathology images," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham, Switzerland: Springer, 2017, pp. 83–91.

[15] S. Valverde, M. Salem, M. Cabezas, D. Pareto, J. C. Vilanova, L. Ramió-Torrentà, À. Rovira, J. Salvi, A. Oliver, and X. Lladó, "One-shot domain adaptation in multiple sclerosis lesion segmentation using convolutional neural networks," 2018, *arXiv:1805.12415*. [Online]. Available: http://arxiv.org/abs/1805.12415

[16] G. Wang, W. Li, M. A. Zuluaga, R. Pratt, P. A. Patel, M. Aertsen, T. Doel, A. L. David, J. Deprest, S. Ourselin, and T. Vercauteren, "Interactive medical image segmentation using deep learning with image-specific fine tuning," *IEEE Trans. Med. Imag.*, vol. 37, no. 7, pp. 1562–1573, Jul. 2018.

[17] R. Huang, J. A. Noble, and A. I. Namburete, "Omni-supervised learning: Scaling up to large unlabelled medical datasets," in *Proc. MICCAI*. Cham, Switzerland: Springer, 2018, pp. 572–580.

[18] C. Rosenberg, M. Hebert, and H. Schneiderman, "Semi-supervised self-training of object detection models," in *Proc. 7th IEEE Workshops Appl. Comput. Vis. (WACV/MOTION)*, Jan. 2005, pp. 29–36.

[19] X. Zhan, Z. Liu, P. Luo, X. Tang, and C. C. Loy, "Mix-and-match tuning for self-supervised semantic segmentation," 2017, *arXiv:1712.00661*. [Online]. Available: http://arxiv.org/abs/1712.00661

[20] J. Wang, Y. Cheng, C. Guo, Y. Wang, and S. Tamura, "Shape–intensity prior level set combining probabilistic atlas and probability map constrains for automatic liver segmentation from abdominal CT images," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 11, no. 5, pp. 817–826, May 2016.

[21] C. Shi, J. Wang, and Y. Cheng, "Sparse representation-based deformation model for atlas-based segmentation of liver CT images," in *Proc. Int. Conf. Image Graph.* Cham, Switzerland: Springer, 2015, pp. 410–419.

[22] C. Shi, Y. Cheng, F. Liu, Y. Wang, J. Bai, and S. Tamura, "A hierarchical local region-based sparse shape composition for liver segmentation in CT scans," *Pattern Recognit.*, vol. 50, pp. 88–106, Feb. 2016.

[23] C. Shi, Y. Cheng, J. Wang, Y. Wang, K. Mori, and S. Tamura, "Low-rank and sparse decomposition based shape model and probabilistic atlas for automatic pathological organ segmentation," *Med. Image Anal.*, vol. 38, pp. 30–49, May 2017.

[24] Y. Cheng, X. Hu, J. Wang, Y. Wang, and S. Tamura, "Accurate vessel segmentation with constrained B-snake," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2440–2455, Aug. 2015.

[25] Y. Zheng, "Cross-modality medical image detection and segmentation by transfer learning of shapel priors," in *Proc. IEEE 12th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2015, pp. 424–427.

[26] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Comput. Vis. Image Understand.*, vol. 61, no. 1, pp. 38–59, Jan. 1995.

[27] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*. Cham, Switzerland: Springer, 2015, pp. 234–241.

[28] X. Yang, C. Bian, L. Yu, D. Ni, and P.-A. Heng, "Hybrid loss guided convolutional networks for whole heart parsing," in *Proc. STACOM*. Cham, Switzerland: Springer, 2017, pp. 215–223.

[29] Q. Dou, L. Yu, H. Chen, Y. Jin, X. Yang, J. Qin, and P.-A. Heng, "3D deeply supervised network for automated segmentation of volumetric medical images," *Med. Image Anal.*, vol. 41, pp. 40–54, Oct. 2017.

[30] Y. Wang, N. Wang, M. Xu, J. Yu, C. Qin, X. Luo, X. Yang, T. Wang, A. Li, and D. Ni, "Deeply-supervised networks with threshold loss for cancer detection in automated breast ultrasound," *IEEE Trans. Med. Imag.*, vol. 39, no. 4, pp. 866–876, Apr. 2020.

[31] K. Chitta, J. Feng, and M. Hebert, "Adaptive semantic segmentation with a strategic curriculum of proxy labels," 2018, *arXiv:1811.03542*. [Online]. Available: http://arxiv.org/abs/1811.03542

[32] D. Shen, Y. Zhan, and C. Davatzikos, "Segmentation of prostate boundaries from ultrasound images using statistical shape model," *IEEE Trans. Med. Imag.*, vol. 22, no. 4, pp. 539–551, Apr. 2003.

[33] X. Zhuang, K. S. Rhode, R. S. Razavi, D. J. Hawkes, and S. Ourselin, "A registration-based propagation framework for automatic whole heart segmentation of cardiac MRI," *IEEE Trans. Med. Imag.*, vol. 29, no. 9, pp. 1612–1625, Sep. 2010.

[34] M. R. Avendi, A. Kheradvar, and H. Jafarkhani, "A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac MRI," *Med. Image Anal.*, vol. 30, pp. 108–119, May 2016.

[35] J. Duan, G. Bello, J. Schlemper, W. Bai, T. J W Dawes, C. Biffi, A. de Marvao, G. Doumou, D. P O'Regan, and D. Rueckert, "Automatic 3D bi-ventricular segmentation of cardiac images by a shape-refined multi-task deep learning approach," 2018, *arXiv:1808.08578*. [Online]. Available: http://arxiv.org/abs/1808.08578

[36] C. Zotti, Z. Luo, O. Humbert, A. Lalande, and P.-M. Jodoin, "Gridnet with automatic shape prior registration for automatic mri cardiac segmentation," in *Proc. STACOM*. Cham, Switzerland: Springer, 2017, pp. 73–81.

[37] H. Ravishankar, R. Venkataramani, S. Thiruvenkadam, P. Sudhakar, and V. Vaidya, "Learning and incorporating shape models for semantic segmentation," in *Proc. MICCAI*. Cham, Switzerland: Springer, 2017, pp. 203–211.

[38] H.-H. Chang, A. H. Zhuang, D. J. Valentino, and W.-C. Chu, "Performance measure characterization for evaluating NeuroImage segmentation algorithms," *NeuroImage*, vol. 47, no. 1, pp. 122–135, Aug. 2009.

[39] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," 2017, *arXiv:1703.10593*. [Online]. Available: http://arxiv.org/abs/1703.10593

[40] L. Wang, H.-M. Lai, G. J. Barker, D. H. Miller, and P. S. Tofts, "Correction for variations in MRI scanner sensitivity in brain studies with histogram matching," *Magn. Reson. Med.*, vol. 39, no. 2, pp. 322–327, Feb. 1998.

[41] L. Wu, Y. Xin, S. Li, T. Wang, P.-A. Heng, and D. Ni, "Cascaded fully convolutional networks for automatic prenatal ultrasound image segmentation," in *Proc. IEEE 14th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2017, pp. 663–666.

[42] S. Candemir, S. Jaeger, K. Palaniappan, J. P. Musco, R. K. Singh, Z. Xue, A. Karargyris, S. Antani, G. Thoma, and C. J. McDonald, "Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration," *IEEE Trans. Med. Imag.*, vol. 33, no. 2, pp. 577–590, Feb. 2014.

[43] M. Campello and K. Lekadir, "Multi-centre, multi-vendor & multi-disease cardiac image segmentation challenge (m&ms)," in *Medical Image Computing and Computer Assisted Intervention*. 2020.

[44] O. Bernard *et al.*, "Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved?" *IEEE Trans. Med. Imag.*, vol. 37, no. 11, pp. 2514–2525, Nov. 2018.

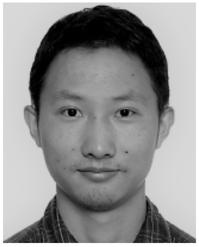
[45] Z. Wang and Z. Wang, "Fully automated segmentation of the left ventricle in magnetic resonance images," 2020, *arXiv:2007.10665*. [Online]. Available: http://arxiv.org/abs/2007.10665

[46] Z. Wang, "Automatic and optimal segmentation of the left ventricle in cardiac magnetic resonance images independent of the training sets," *IET Image Process.*, vol. 13, no. 10, pp. 1725–1735, Aug. 2019.

[47] X. Zhuang and J. Shen, "Multi-scale patch and multi-modality atlases for whole heart segmentation of MRI," *Med. Image Anal.*, vol. 31, pp. 77–87, Jul. 2016.

CHAOYU CHEN received the B.Eng. degree from Southern Medical University, in 2018. He is currently pursuing the master's degree with the Medical Ultrasound Image Computing (MUSIC) Laboratory, Health Science Center, School of Biomedical Engineering, Shenzhen University.

His current research interest includes deep learning in medical image analysis.



XIN YANG received the master's degree in biomedical engineering from Shenzhen University, Shenzhen, China, in 2015, and the Ph.D. degree in computer science and engineering from The Chinese University of Hong Kong, Hong Kong, in 2019.

He is currently an Assistant Professor with the Health Science Center, School of Biomedical Engineering, Shenzhen University. His research interests include ultrasound image analysis, cardiac image analysis, and computer graphics.

HAORAN DOU received the B.Eng. degree from Sichuan University, in 2017. He is currently pursuing the master's degree with the Medical Ultrasound Image Computing (MUSIC) Laboratory, Health Science Center, School of Biomedical Engineering, Shenzhen University.

His current research interest includes medical image segmentation.

RUOBING HUANG received the bachelor's degree from the Beijing Institute of Technology, China, in 2014, and the D.Phil. degree from Oxford University, U.K., in 2018, under the supervision of Prof. A. Noble.

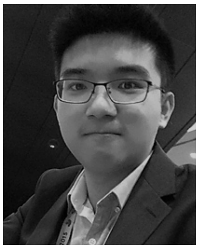
She continued her research as a Postdoctoral Research Fellow of Oxford University until August 2019. She is currently an Assistant Professor with Shenzhen University, China. Her research interests include deep learning, ultrasound image analysis, and computer-assisted diagnosis.

XIAOQIONG HUANG received the B.Eng. degree from the Guangdong University of Technology. She is currently pursuing the master's degree with the Medical Ultrasound Image Computing (MUSIC) Laboratory, Health Science Center, School of Biomedical Engineering, Shenzhen University.

Her current research interest includes deep learning in domain adaption.

XU WANG received the master's degree from Shenzhen University.

In 2016, he joined the Medical Ultrasound Image Computing (MUSIC) Laboratory, Health Science Center, School of Biomedical Engineering, Shenzhen University. His research interests include deep learning and medical image processing.

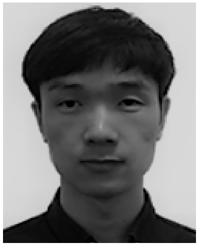


CHONG DUAN received the bachelor's degree in chemistry from Nankai University, Tianjin, China, in 2012, and the PhD degree in chemistry from Washington University, St. Louis, MO, USA, in 2017.

He is currently an Imaging Lead with the Department of Early Clinical Development, Pfizer's Worldwide Research and Development, Pfizer Inc. His research interests include medical imaging physics and analytics.

SHENGLI LI received the master's degree in radiology from the Xiangya School of Medicine, Changsha, China, in 1994.

He is currently a Chief Physician and a Professor with the Department of Ultrasound, Affiliated Shenzhen Maternal and Child Healthcare Hospital, Nanfang Medical University, Guangzhou, China. His current research interest includes ultrasound diagnosis.



WUFENG XUE (Member, IEEE) received the bachelor's degree in automation and the Ph.D. degree in signal and information processing from Xi'an Jiaotong University, China, in 2009 and 2016, respectively.

From 2016 to 2018, he was a Postdoctoral Research Fellow of the University of Western Ontario, Canada. Since 2018, he has been an Associate Professor with the School of Biomedical Engineering, Shenzhen University. He was awarded the Young Scientist Award from the MICCAI in 2017. His research interests include medical image analysis, pattern recognition, and machine learning.

PHENG ANN HENG (Senior Member, IEEE) received the Ph.D. degree in computer science from Indiana University, Bloomington, IN, USA.

He is currently a Professor with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, where he is also the Director of the Virtual Reality, Visualization, and Imaging Research Centre. He is also the Director of the Research Center for Human–Computer Interaction, Shenzhen Institute of Advanced Integration Technology, Chinese Academy of Sciences, Shenzhen, China. His research interests include VR and AI applications in medicine, visualization, medical imaging, human–computer interfaces, and interactive graphics.

DONG NI (Member, IEEE) received the bachelor's and master's degrees in biomedical engineering from Southeast University, Nanjing, China, in 2000 and 2003, respectively, and the Ph.D. degree in computer science and engineering from The Chinese University of Hong Kong, Hong Kong, in 2009.

From 2009 to 2010, he was a Postdoctoral Fellow of the School of Medicine, University of North Carolina at Chapel Hill, USA. Since 2010, he has been with Shenzhen University, China, where he is currently a Professor and the Associate Dean of the Health Science Center, School of Biomedical Engineering. He founded the Medical Ultrasound Image Computing (MUSIC) Laboratory, Shenzhen University. His research interests include ultrasound image analysis, image guided surgery, and pattern recognition.