Contrastive Semi-Supervised Learning for Domain Adaptive Segmentation Across Similar Anatomical Structures

Ran Gu, Jingyang Zhang, Guotai Wang, Wenhui Lei, Tao Song, Xiaofan Zhang, Kang Li, and Shaoting Zhang

***Abstract*— Convolutional Neural Networks (CNNs) have achieved state-of-the-art performance for medical image segmentation, yet need plenty of manual annotations for training. Semi-Supervised Learning (SSL) methods are promising to reduce the requirement of annotations, but their performance is still limited when the dataset size and the number of annotated images are small. Leveraging existing annotated datasets with similar anatomical structures to assist training has a potential for improving the model's performance. However, it is further challenged by the cross-anatomy domain shift due to the image modalities and even different organs in the target domain. To solve this problem, we propose Contrastive Semi-supervised learning for Cross Anatomy Domain Adaptation (CS-CADA) that adapts a model to segment similar structures in a target domain, which requires only limited annotations in the target domain by leveraging a set of existing annotated images of similar structures in a source domain. We use Domain-Specific Batch Normalization (DSBN) to individually normalize feature maps for the two anatomical domains, and propose a cross-domain contrastive learning strategy to encourage extracting domain invariant features. They are integrated into a Self-Ensembling Mean-Teacher (SE-MT) framework to exploit unlabeled target domain images with a prediction consistency constraint. Extensive experiments show that our CS-CADA is able to solve the challenging cross-anatomy domain shift problem, achieving accurate segmentation of coronary arteries in X-ray images with the help of retinal vessel images and cardiac MR images with the help of fundus images, respectively, given only a small number of annotations in the target domain. Our code is available at https://github.com/HiLab-git/DAG4MIA.**

***Index Terms*— Semi-supervised learning, cross-anatomy domain adaptation, contrastive learning.**

Manuscript received 20 June 2022; revised 10 August 2022 and 17 September 2022; accepted 21 September 2022. Date of publication 26 September 2022; date of current version 29 December 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 61901084 and Grant 81771921; and in part by the Department of Science and Technology of Sichuan Province, China, under Grant 20ZDYF2817. *(Ran Gu and Jingyang Zhang contributed equally to this work.) (Corresponding authors: Guotai Wang; Shaoting Zhang.)*

Ran Gu and Guotai Wang are with the School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China, and also with the Shanghai AI Laboratory, Shanghai 200240, China (e-mail: guran924@std.uestc.edu.cn; guotai.wang@uestc.edu.cn).

Jingyang Zhang is with the School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China, and also with the School of Biomedical Engineering, ShanghaiTech University, Shanghai 200240, China (e-mail: zjysjtu1994@gmail.com).

Wenhui Lei and Xiaofan Zhang are with the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China, and also with the Shanghai AI Laboratory, Shanghai 200240, China (e-mail: wenhui.lei@sjtu.edu.cn; xiaofan.zhang@sjtu.edu.cn).

Tao Song is with SenseTime Research, Shanghai 200240, China (e-mail: songtao@sensetime.com).

Kang Li is with the West China Hospital, Sichuan University, Chengdu 611731, China (e-mail: likang@wchscu.cn).

Shaoting Zhang is with the School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China, also with the Shanghai AI Laboratory, Shanghai 200240, China, and also with SenseTime Research, Shanghai 200240, China (e-mail: zhangshaoting@uestc.edu.cn).

Digital Object Identifier 10.1109/TMI.2022.3209798

I. Introduction

RECENTLY, Convolutional Neural Networks (CNNs) have achieved remarkable progress in medical image segmentation [1], [2], yet requiring a large amount of manual annotations for training images, which is highly time-consuming and labor-intensive to collect. Therefore, it is desired to reduce the manual annotations for model training while maintaining the segmentation performance. Semi-Supervised Learning (SSL) has been widely used to reduce the required annotations, as it only requires a small set of labeled data with the availability of a large set of unlabeled data [3]. Although SSL methods have achieved promising performance in medical image segmentation [4], [5], they often rely on a very large set of unannotated images and the performance is still limited when the number of labeled images is very small. For a lot of medical imaging applications, it is not only time-consuming to acquire annotations, but also difficult and expensive to collect a large set of unannotated images, leading to a small set of available training samples [3]. In such cases, it is challenging for most existing SSL methods to achieve high performance with a small training set of which part samples are annotated.

Although the small set of training samples as well as lack of annotations for a given task are common problems in the field of medical image segmentation, there are many existing datasets with full annotations. It is desirable to leverage such datasets to assist the training of a model for a target

task, i.e., adapting a model trained with existing datasets (i.e., source domain) to a specific target dataset (i.e., target domain). As the source and target domain images are often acquired with different imaging protocols, such as different modalities, contrasts, patient groups and even anatomical structures, the performance is limited if such pre-trained models are directly applied to the target domain images for inference. Recently, Domain Adaptation (DA) is attracted increasing attentions, which assumes that the same task is involved in the source and target domains, e.g., adapting a model trained with annotated heart Magnetic Resonance Images (MRI) to segment heart Computed Tomography (CT) images [6]. To alleviate the performance gap between the two domains, a widely used method is to fine-tune the pre-trained models with target domain images [7]. However, the fine-tuning process requires annotations in the target domain and cannot leverage unannotated images for training. Alternatively, Unsupervised Domain Adaptation (UDA) methods have been increasingly investigated to adapt a model trained with a source domain to a target domain. UDA methods do not require annotations in the target domain and usually translate source-domain images [6] to target domain-like images or learn domain-invariant features [8] to achieve good results.

However, most of the state-of-the-art DA methods for medical image segmentation require that the source and target domains have the same set of anatomical structures even they can be from different imaging modalities [6]. Such a requirement prevents these methods from leveraging images of other anatomical structures for training, and unlocking this requirement would enlarge the scope of candidate source domains, which helps to improve the segmentation in the target domain when a source domain with exactly the same anatomical structures is not available. For instance, for segmentation of coronary arteries from 2D X-ray Angiogram (XA) with limited annotations, it is hard to find an annotated dataset with the same structure as the target domain. However, there are a lot of public fundus images with annotated retinal vessels (e.g., DRIVE [9] and STARE [10]), where the retinal vessels share similar tubular structures with the coronary arteries, as shown in fist row of Fig. 1. Another example is the similar circular structures between Left Ventricle blood cavity (LV) and the Myocardium (Myo) in Cardiac MR (CMR) images and the optic cup and disc in public Retinal images (Retinal), as shown in the second row of Fig.1. Hence, it is promising to transfer the knowledge from these retinal vessel datasets to the coronary artery segmentation task [11], as they can be regarded as cost-free source domain images. However, the different morphologies and contexts of these two kinds of vessels make it hard to achieve accurate results for existing UDA methods that are designed to deal with the same anatomical structures.

Unlike previous works, we investigate the cross-anatomy semi-supervised domain adaptation problem for medical image segmentation. Compared with existing domain adaptation methods that require the segmented objects to be the same in the source and target domains, our method relaxes this requirement and enables adapting a model to segment a different anatomical structure to the source domain. Compared with existing semi-supervised methods requiring the training

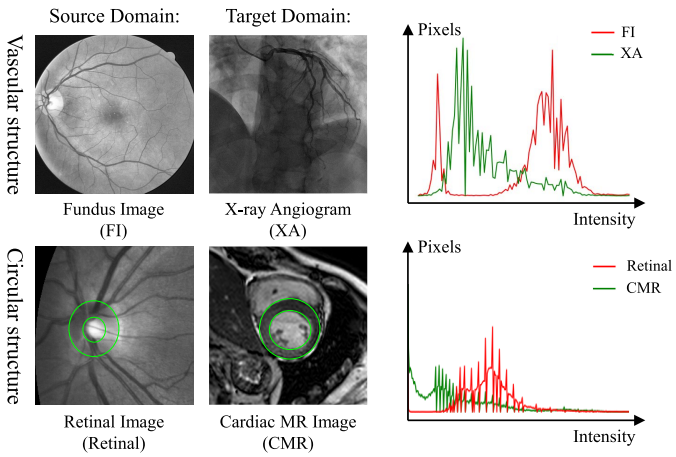

Fig. 1. Illustration of the challenging cross-anatomy domain shift in two scenarios. Row 1 shows the domain shift between fundus image (FI) and X-ray angiogram (XA) with similar vascular structures. Row 2 shows the cross-circular domain shift between Retinal image (Retinal) and Cardiac MR image (CMR) with circular structures. Note the different intensity histograms, morphologies, and contexts of the similar anatomical structures in the source and target domains.

samples from the same domain, our method leverages images from other domains with similar anatomical structures to assist the training process, which can improve the segmentation performance in the target domain.

Our proposed method is referred to as Contrastive Semi-supervised learning for Cross Anatomy Domain Adaptation (CS-CADA) for medical image segmentation. The main contributions are:

1) To reduce the annotation cost and overcome the problem of limited training images in a target domain, we propose to leverage an existing annotated dataset of a similar anatomical structure from a different domain for training. A novel framework called CS-CADA that is a generalization of existing semi-supervised and domain adaptation methods is introduced.

2) To deal with domain shift between the two domains while transferring the knowledge of similar anatomical structures, we use Domain-Specific Batch Normalization (DSBN) [12] with shared convolutional kernels for the two domains, and integrate it into a Mean Teacher (MT)-based consistency regularization framework to leverage unannotated images in the target domain.

3) To better learn domain-invariant features, we propose a cross-domain contrastive learning strategy, where a pair of features from the two domains based on their individual DSBN are treated as positive pairs and otherwise negative pairs.

4) Extensive experiments on two different scenarios (coronary artery and cardiac MRI segmentation with the assistance of retinal vessel images and fundus images, respectively) demonstrate that our proposed CS-CADA achieved accurate segmentation results with better performance than existing semi-supervised and domain adaptation methods when annotations in the target domain are limited.

A preliminary version of this work was published in 2021 [13], where we applied the combination of DSBN and SE-MT for coronary artery segmentation. In this extension, we provide detailed descriptions of our framework, and

introduce the cross-domain contrastive learning strategy for better performance. The method has also been validated with more extensive experiments of different applications in this work.

II. Related Works

A. Semi-Supervised Learning

Semi-supervised learning methods have been widely used to reduce the amount of required annotations for medical image segmentation [3]. Existing methods mainly use strategies such as pseudo label [14], adversarial training [15] and consistency regularization [4]. Pseudo label-based methods train a model on annotated images to generate masks for unannotated images that are then used to update the segmentation model. To improve the quality of pseudo labels, different strategies including randomly selected propagation [16] and uncertainty-based refinement [17] have been proposed. Adversarial learning-based methods [15] use discriminators to encourage the predictions of unannotated images to be similar to those of the annotated ones, where prior information [18] may be used to improve the performance. Consistency regularization methods encourage predictions from one input image under different perturbations to be consistent, e.g., transformation consistency [19], dual-task consistency [20], and perturbation-based consistency [21].

Mean teacher is a widely used consistency-based SSL method [5]. It uses a self-ensembling of a student model as the teacher model, and encourages consistent predictions between the two models [22]. It has been extended to Uncertainty Aware Mean-Teacher (UA-MT) [21] and combined with transformation consistency to achieve better results [23] for medical image segmentation. However, most existing SSL methods assume the training samples are from the same domain and hardly obtain high performance when the available images and annotations are limited. In this work, we leverage annotated images from an existing dataset with similar structures (e.g., a public dataset) to assist the semi-supervised training for better performance.

B. Transfer Learning and Domain Adaptation

Transfer Learning (TL) aims to transfer knowledge learned from a source dataset to deal with data in a new target dataset [24]. Early transfer learning methods mainly fine-tune part or the whole set of parameters in a pre-trained model with an annotated target dataset [25], [26]. Chen *et al.* [27] fine-tuned a pre-trained CNN for localizing standard planes in ultrasound images. Tajbakhsh *et al.* [26] demonstrated that knowledge can be transferred from natural images to medical images through fine-tuning. Although fine-tuning is easy to implement straightforward for transfer learning, it is always faced with three main issues, i.e., where, how and when to fine-tune [24]. If the source and target datasets have a large gap, the fine-tuning is less effective. Differently from classical transfer learning, Domain Adaptation (DA) deals with the same set of target objects in different domains (e.g., imaging protocols, patient groups, or intensity distribution), which is more effective to transfer the knowledge in the source domain to the target domain due to the shared anatomical structure. In addition, it can leverage the source and target images simultaneously for training, rather than using the source and target images in two independent stages in typical transfer learning.

Domain Adaptation has been widely used to alleviate the performance degradation when the distribution of target data differs from that of source data [28], [29]. DA methods mainly have three categorises based on the type of annotations. The first is fully-supervised DA, where fine-tuning is the most representative technique for adapting a trained model to the target domain with full annotations [25], [30]. The second is weakly or semi-supervised DA where only coarse or partial annotations in the target domain are used for model adaptation [31]. For example, Dorent *et al.* [32] employed scribbles in the target domain to perform model adaptation for vestibular schwannoma segmentation. Li *et al.* [33] used a dual-teacher semi-supervised domain adaptation method when a small ratio of the target domain images have annotations. Additionally, Unsupervised Domain Adaptation (UDA) does not require annotations in the target domain. UDA methods usually use Generative Adversarial Networks (GAN) to achieve image-level or feature-level alignment between the source and target domains. For example, Zhu *et al.* [34] used Cycle-GAN to convert source domain images to target-domain like images to reduce the domain gap. Kamnitsas *et al.* [35] and Dou *et al.* [8] used GAN to obtain domain invariant features for adaptation. The Simultaneous Image and Feature Alignment (SIFA) [6] combined the advantage of these two categories and has achieved state-of-the-art performance for UDA task.

Although existing annotation-efficient DA methods achieved promising performance in the target domain with limited annotations, most of them can only deal with two domains with the same anatomical structure, and are not applicable for cross-anatomy domain adaptation. In this work, we deal with the domain shift between two datasets of different anatomical structures with similar shapes.

C. Contrastive Learning

Contrastive learning commonly employs a contrastive loss to enforce representations to be similar for positive pairs and dissimilar for negative pairs [36]. Previous contrastive learning methods are mainly proposed as a self-supervised pre-training strategy to train a powerful and representational feature extractor that can be fine-tuned for down-stream tasks [37], [38].

Recently, contrastive learning has also been used for domain adaptation. Kang *et al.* [39] proposed an end-to-end contrasitve adaptation network that minimizes the intra-class domain discrepancy and maximizes the inter-class discrepancy. Singh *et al.* [40] employed class-wise and instance-level contrastive learning to respectively minimize the inter-domain and inter-domain discrepancy in semi-supervised domain adaptation. In this work, we design a novel contrastive learning strategy for cross-anatomy domain adaptation, which encourages the model to extract comprehensive domain-invariant features across similar anatomical structures.

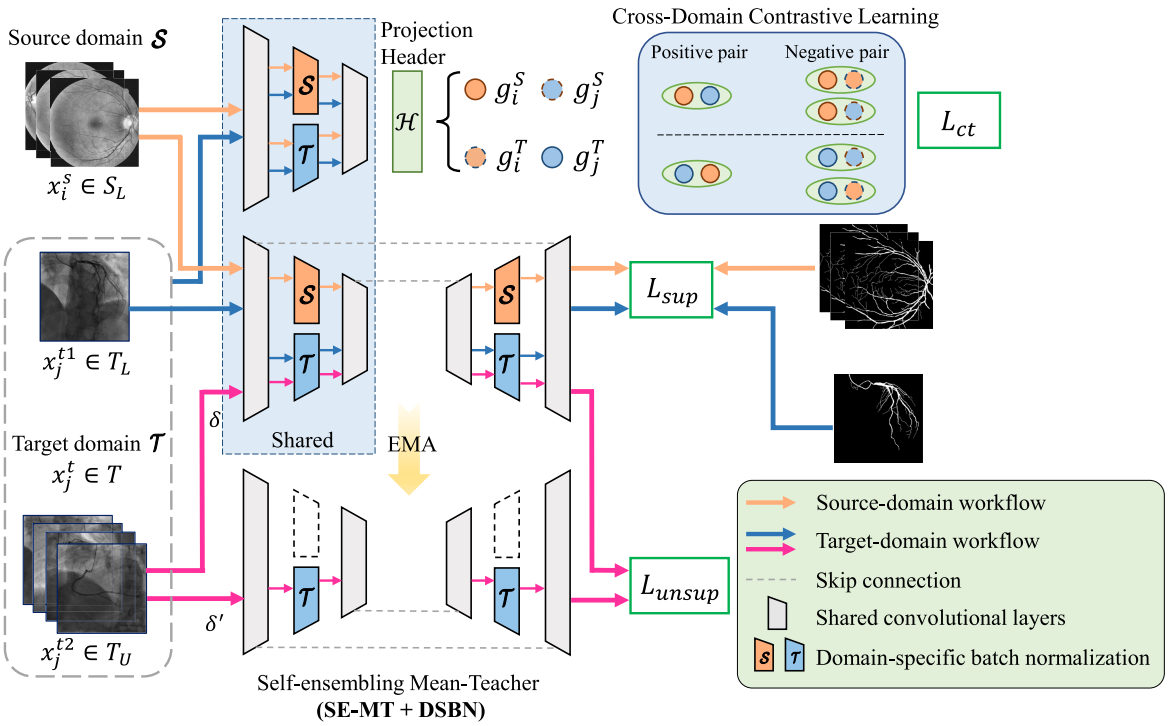

Fig. 2. The flowchart of the proposed CS-CADA that consists of three parts: 1) a segmentation network with Domain-Specific Batch Normalization (DSBN); 2) a Self-ensembling Mean-Teacher (SE-MT) architecture; and 3) a cross-domain contrastive learning block. $g_i^{\mathcal{S}}$ and $g_i^{\mathcal{T}}$ are features of source-domain image x_i^s normalized by the source- and target-specific BN layers, respectively. $g_j^{\mathcal{S}}$ and $g_j^{\mathcal{T}}$ are features of target-domain image x_j^t normalized by the source- and target-specific BN layers, respectively.

III. Methods

Let $\mathcal{S}$ and $\mathcal{T}$ denote the source and the target domains, respectively. We denote the existing annotated source domain dataset as $S_L = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}\}$, and use $T_L = \{(x_j^{t1}, y_j^{t1})\}_{j=1}^{N_{t1}}\}$ to denote a small set of annotated images in the target domain. Additionally, another set of unannotated images in the target domain is denoted as $T_U = \{(x_j^{t2})\}_{j=1}^{N_{t2}}$, where N_s, N_{t1} and N_{t2} are the number of samples in the three datasets, respectively. The proposed CS-CADA is depicted in Fig. 2, which consists of three parts: 1) a segmentation network with Domain-Specific Batch Normalization (DSBN) that learns from S_L and T_L to provide a supervision guidance to bridge the cross-anatomy discrepancy; and 2) a Self-Ensembling Mean-Teacher (SE-MT) that imposes an unsupervised consistency on T_U to further enhance data efficiency, and 3) a contrastive learning that encourages the model to capture domain-invariant features. Without loss of generality, we adopt the classical U-Net [41] as the backbone for segmentation.

A. Joint Learning With Domain-Specific Batch Normalization (DSBN)

Considering the intensity distribution shift between source and target domains, directly taking S_L and T_L for training without dealing with their difference would have limited performance, since the model will be misguided by statistical variations between domain $\mathcal{S}$ and $\mathcal{T}$ and thus fail to lean general feature representations from the two domains.

To solve the problem, we introduce DSBN block to the network that consists of two types of Batch Normalization (BN) and each of them is in charge of one domain to effectively tackle the inter-domain discrepancy [42]. In our method, DSBN adopts respective BN parameters for domain $\mathcal{S}$ and $\mathcal{T}$ due to the cross-anatomical structure discrepancy. Meanwhile, convolutional kernels are shared across domain $\mathcal{S}$ and $\mathcal{T}$ to learn general representations for the similar anatomical structures. Formally, let $f^d \in \mathbb{R}^{N\times H\times W}$ denote feature maps at each channel given by an input from domain $d \in \{\mathcal{S}, \mathcal{T}\}$. The DSBN normalizes each feature respectively and then applies affine transformation with trainable parameters that are specific to the certain domain d, i.e., re-scale parameters γ^d and bias parameters β^d:

$$\hat{f}^d = \gamma^d \cdot \overline{f}^d + \beta^d, \quad \text{where} \quad \overline{f}^d = \frac{f^d - \mu^d}{\sqrt{(\sigma^d)^2 + \varepsilon}}, \tag{1}$$

where $\hat{f}^d$ is the DSBN output. μ^d and σ^d are the mean and standard deviation of the features within a mini-batch containing N samples, and ε is a small number for numeric stability.

Let θ_{en} and θ_{de} represent the shared convolutional parameters in the encoder and decoder of the model, respectively. $\{\gamma^d, \beta^d\}$ represents a set of trainable parameters in DSBN at domain d. Formally, the parameter set for the source domain can be summarized as $\Theta^{\mathcal{S}} = [\theta_{en}, \theta_{de}, \gamma^{\mathcal{S}}, \beta^{\mathcal{S}}]$, and that for the target domain is $\Theta^{\mathcal{T}} = [\theta_{en}, \theta_{de}, \gamma^{\mathcal{T}}, \beta^{\mathcal{T}}]$. DSBN supplies domain-specific variables to handle domain-specific distributions and maps stylistic features to a common space by performing individual feature normalization, which can effectively alleviate the inter-domain discrepancy [42]. During

training, DSBN calculates the mean and standard deviation of features for each domain separately, i.e., $\overline{\mu}^d$ and $\overline{\sigma}^d$. In the test phase, the estimated $\overline{\mu}^d$ and $\overline{\sigma}^d$ from a moving average in the training stage for each domain are used for whitening input activation.

Given image-annotation pairs $(x_i^s, y_i^s) \in S_L$ from domain $\mathcal{S}$ and $(x_j^{t1}, y_j^{t1}) \in T_L$ from domain $\mathcal{T}$, we define a supervised loss function to jointly optimize these parameter sets:

$$L_{sup} = \sum_{i=1}^{N_s} L_{seg}(p_i^s, y_i^s) + \sum_{j=1}^{N_{t1}} L_{seg}(p_j^{t1}, y_j^{t1}), \tag{2}$$

where $p_i^s = \psi(x_i^s; \Theta^{\mathcal{S}})$ and $p_j^{t1} = \psi(x_j^{t1}; \Theta^{\mathcal{T}})$ are predictions of x_i^s and x_j^{t1} using the corresponding parameter sets $\Theta^{\mathcal{S}}$ and $\Theta^{\mathcal{T}}$, respectively. L_{seg} denotes a hybrid segmentation loss that consists of the cross-entropy loss and Dice loss.

B. Self-Ensembling Mean Teacher (SE-MT) With DSBN

Although the introduced DSBN enables a model to learn from annotated S_L and T_L, the small-scale T_L still limits the performance of the model. To deal with this problem, we employ a Self-Ensembling Mean Teacher (SE-MT) architecture to exploit unannotated images T_U in the target domain. Specifically, the teacher model $\Theta^{\mathcal{T}'}$ is defined as an Exponential Moving Average (EMA) of the student model in the target domain, its parameter in the k-th training step is:

$$\Theta_k^{\mathcal{T}'} = \alpha \Theta_{k-1}^{\mathcal{T}'} + (1-\alpha)\Theta_k^{\mathcal{T}} \tag{3}$$

where $\alpha \in [0, 1]$ is the EMA decay rate [43]. The teacher model will guide the student model to give more reliable predictions for unannotated images. Finally, we define an unsupervised consistency loss (L_{unsup}) between the predictions of student model and teacher model for the same input $x_j^{t2} \in T_U$ with different random perturbations δ and δ':

$$L_{unsup} = \sum_{j=1}^{N_{t2}} L_{mse}\left(\psi(x_j^{t2}; \Theta^{\mathcal{T}}, \delta), \psi(x_j^{t2}; \Theta^{\mathcal{T}'}, \delta')\right) \tag{4}$$

where $\psi(x_j^{t2}; \Theta^{\mathcal{T}}, \delta)$ and $\psi(x_j^{t2}; \Theta^{\mathcal{T}'}, \delta')$ are predictions given by the student and teacher models, respectively. L_{mse} is the mean square error loss.

C. Cross-Domain Contrastive Learning

To better deal with the appearance and context shift between the source and target domains, we propose a cross-domain contrastive learning strategy to encourage the model to capture domain-invariant features for the similar anatomical structures while being robust against the different image styles.

For the output of the encoder with DSBN of the segmentation network, we use a nonlinear projection header $\mathcal{H}$ to obtain a high-level feature representation. For a source domain image x_i^s, its normalized feature representations based on the source-domain BN and target-domain BN are denoted as $g_i^{\mathcal{S}} = f(x_i^s; \theta_{\mathcal{H}}, \theta_{en}, \gamma^{\mathcal{S}}, \beta^{\mathcal{S}})$ and $g_i^{\mathcal{T}} = f(x_i^s; \theta_{\mathcal{H}}, \theta_{en}, \gamma^{\mathcal{T}}, \beta^{\mathcal{T}})$, respectively. Correspondingly, for a target domain image x_j^t, the normalized feature representations based on the source and target domain-specific BNs are denoted as $g_j^{\mathcal{S}} = f(x_j^t; \theta_{\mathcal{H}}, \theta_{en}, \gamma^{\mathcal{S}}, \beta^{\mathcal{S}})$ and $g_j^{\mathcal{T}} = f(x_j^t; \theta_{\mathcal{H}}, \theta_{en}, \gamma^{\mathcal{T}}, \beta^{\mathcal{T}})$, respectively.

As the DSBN aims to normalize the feature maps of two domains respectively so that they are mapped to a common feature distribution space to alleviate the domain gap. Given a pair of images from the source and target domains x_i^s and x_j^t, $g_i^{\mathcal{S}}$ and $g_j^{\mathcal{T}}$ are both normalized features in the common distribution space based on their specific BNs. They should be similar to each other, and thus be set as a positive pair. For the pair of $(g_i^{\mathcal{S}}, g_i^{\mathcal{T}})$, they are resulted from the same feature of a source-domain image x_i^s normalized by two different BNs, they should be different from each other due to the different styles associated with the two BNs. Therefore, they are set as a negative pair.

Following the standard formula of self-supervised contrastive loss [37], [44], we define a source to target domain contrastive loss as:

$$L_{ct}^{s2t} = -\mathbb{E}_{x_i^s, x_j^t \sim \mathcal{S},\mathcal{T}}\left(log\left(\frac{e^{sim(g_i^{\mathcal{S}}, g_j^{\mathcal{T}})/\tau}}{e^{sim(g_i^{\mathcal{S}}, g_j^{\mathcal{T}})/\tau} + \sum_{g \in \mathcal{N}_i} e^{sim(g_i^{\mathcal{S}}, g)/\tau}}\right)\right) \tag{5}$$

where $\mathcal{N}_i = \{g_i^{\mathcal{T}}, g_j^{\mathcal{S}}\}$ are the negative counterparts of $g_i^{\mathcal{S}}$. The $sim(\cdot, \cdot)$ is the cosine similarity between two representations, and $\tau = 0.1$ is the temperature scaling parameter. Correspondingly, for $g_j^{\mathcal{T}}$, we define a target to source domain contrastive loss as:

$$L_{ct}^{t2s} = -\mathbb{E}_{x_i^s, x_j^t \sim \mathcal{S},\mathcal{T}}\left(log\left(\frac{e^{sim(g_j^{\mathcal{T}}, g_i^{\mathcal{S}})/\tau}}{e^{sim(g_j^{\mathcal{T}}, g_i^{\mathcal{S}})/\tau} + \sum_{g \in \mathcal{N}_j} e^{sim(g_j^{\mathcal{T}}, g)/\tau}}\right)\right) \tag{6}$$

where $\mathcal{N}_j = \{g_j^{\mathcal{S}}, g_i^{\mathcal{T}}\}$ are the negative counterparts of $g_j^{\mathcal{T}}$. Finally, the cross-domain contrastive loss is formulated as:

$$L_{ct} = \frac{1}{2}(L_{ct}^{s2t} + L_{ct}^{t2s}) \tag{7}$$

D. Overall Training Loss

The overall loss for training is a combination of the supervised loss L_{sup} in Eq. 2, the unsupervised consistency loss L_{unsup} in Eq. 4 and the contrastive learning loss L_{ct} in Eq. 7. It is formulated as:

$$L = L_{sup} + \lambda_1 \cdot L_{unsup} + \lambda_2 \cdot L_{ct} \tag{8}$$

where λ_1, λ_2 act as trade-off parameters. Once the training process is completed, segmentation results for the target domain images are obtained by a forward propagation on the student model with parameter set $\Theta^{\mathcal{T}}$.

IV. Experimental Results

A. Implementation Details

For experiments, we used the U-Net [41] as segmentation network in the mean-teacher backbone, and extended it with DSBN for domain-specific feature normalization. The EMA decay rate was empirically set to 0.99, and a time-dependent

TABLE I

QUANTITATIVE COMPARISON BETWEEN OUR CS-CADA AND STATE-OF-THE-ART METHODS FOR CARDIAC ARTERY SEGMENTATION. ONLY 10% OF TRAINING IMAGES IN THE TARGET DOMAIN ARE ANNOTATED. FINE-TUNING (LAST) MEANS ONLY UPDATING THE LAST CONVOLUTIONAL BLOCK OF THE DECODER. FINE-TUNING (ALL) MEANS UPDATING ALL THE PARAMETERS OF THE MODEL

Methods		Training set S_L	Training set T_U	Training set T_L	Recall (%)	Precision (%)	Dice (%)
Baseline (source)		✓			34.68±6.22	40.84±2.79	37.31±4.90
Baseline (target)				✓	75.12±4.83	65.87±8.47	69.81±5.35
UDA	ADDA [29]	✓	✓		43.13±15.08	39.49±6.49	40.24±9.34
UDA	SIFA [6]	✓	✓		62.62±10.15	46.66±7.43	52.44±4.62
UDA	SC-GAN [11]	✓	✓		80.87±7.69	57.88±8.73	66.92±6.83
SDA	Fine-tuning(last) [7]	✓		✓	56.76±7.12	52.55±14.63	53.75±11.27
SDA	Fine-tuning(all)	✓		✓	77.44±4.46	73.81±2.75	75.50±2.72
SDA	Joint Training	✓		✓	81.14±6.75	62.66±9.35	70.37±7.37
SDA	X-shape [46]	✓		✓	83.10±4.11	65.11±10.92	72.49±7.69
SDA	DSBN [43]	✓		✓	80.66±4.58	69.12±9.53	74.12±6.62
SSL	SE-MT [47]		✓	✓	82.55±3.89	72.14±8.40	76.70±5.43
SSL	UA-MT [21]		✓	✓	82.40±2.88	69.13±4.35	75.13±3.35
SSL	CPS [14]		✓	✓	82.13±3.36	72.02±3.80	76.69±3.10
SSDA	Dual-T [34]	✓	✓	✓	82.63±2.89	70.33±3.14	75.94±2.53
CS-CADA (ours)		✓	✓	✓	**83.13±3.87**	**77.32±6.55**	**79.28±3.67**

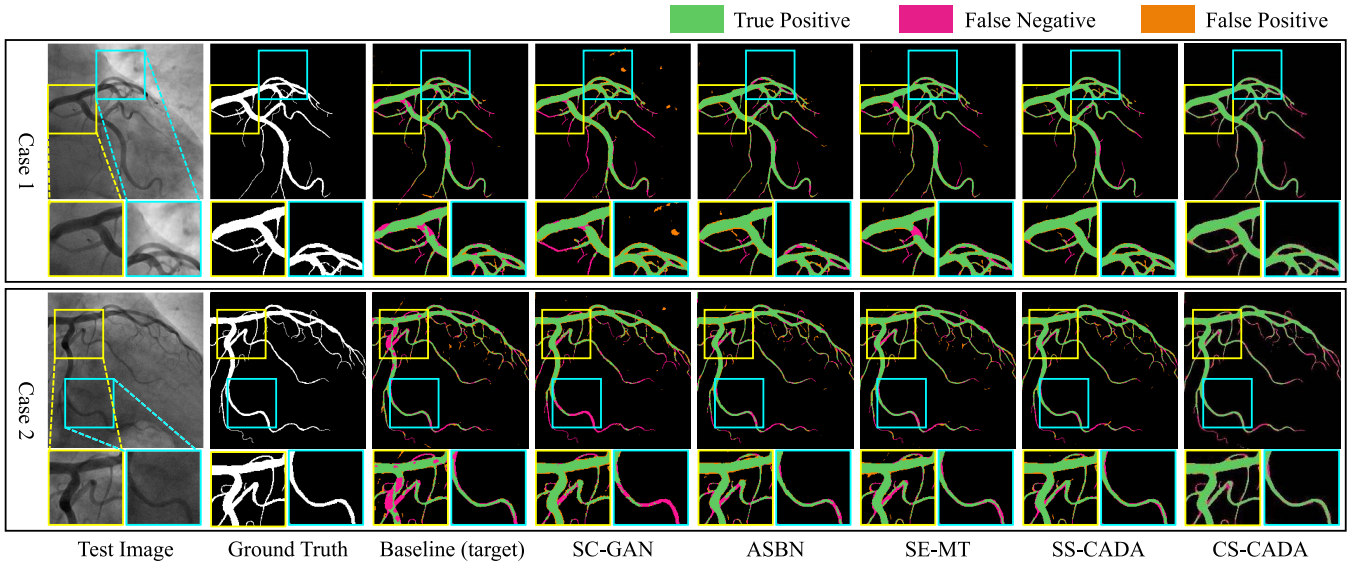

Fig. 3. Visual comparison of different methods for coronary artery segmentation from XAs. The true positives, false negatives and false positives are colored in green, red and orange, respectively. The zoomed views are appended below each case to highlight the segmentation details.

Gaussian warming up function $\alpha(k) = 0.1e^{(-5(1-k/k_{max})^2)}$ was used to dynamically update hyper parameter α, where k denotes the current training iteration and k_{max} is the last iteration. Each model was trained by the Adam optimizer with max iteration $20,000$ and initial learning rate $5e^{-4}$ that was decayed exponentially with power 0.95. For training, the loss function weights are set as $\lambda_1 = 1$ and $\lambda_2 = 0.1$, respectively. The batch size was 8 and 12 for vascular structure segmentation and circular structure segmentation, respectively. And in each batch, half of the images are labeled and half of them are unlabeled. All experiments were implemented by Pytorch with one NVIDIA Geforce GTX 1080 Ti GPU.

Our CS-CADA was validated with two applications: 1) adapting a model from digital retinal images with annotated vessels to segment coronary artery from XAs, and 2) adapting a model from retinal fundus images with annotated optic cup and disk to segment left ventricle blood cavity and myocardium from cardiac MRI. Each of two domain images were processed by contrast limited adaptive histogram equalization and gamma correction.

B. Vascular Structure Segmentation

1) Datasets: The first scenario is modal adaptation for vascular structure segmentation. We used the DRIVE [9] dataset as the source domain dataset S_L. It contains 40 Fundus Images (FI) of retinal vessels acquired from a Canon CR5 nonmydriatic 3CCD camera at 45° field of view. In addition, we collected 191 XAs of 30 patients using a Philips UNIQ FD10 C-arm system with coronary arteries as the target domain. An expert radiologist randomly annotated 14 XAs

TABLE II
QUANTITATIVE ABLATION STUDY OF OUR CS-CADA FOR CORONARY ARTERY SEGMENTATION. SE-MT IS SEMI-SUPERVISED LEARNING ONLY USING T_L AND T_U. DSBN MEANS USING THE DOMAIN-SPECIFIC BATCH NORMALIZATION FOR LEARNING FROM S_L AND T_L. SS-CADA MEANS LEARNING FROM S_L, T_L AND T_U AT THE SAME TIME BASED ON OUR INTRODUCED DSBN AND SE-MT, WITHOUT USING CONTRASTIVE LEARNING

Methods	Recall (%)	Precision (%)	Dice (%)
Baseline (target)	75.12±4.83	65.87±8.47	69.81±5.35
SE-MT	82.55±3.89	72.14±8.40	76.70±5.43
DSBN	80.66±4.58	69.12±9.53	74.12±6.62
SS-CADA [13]	**83.27±3.95**	75.12±6.59	78.84±4.60
CS-CADA (ours)	83.13±3.87	**77.32±6.55**	**79.28±3.67**

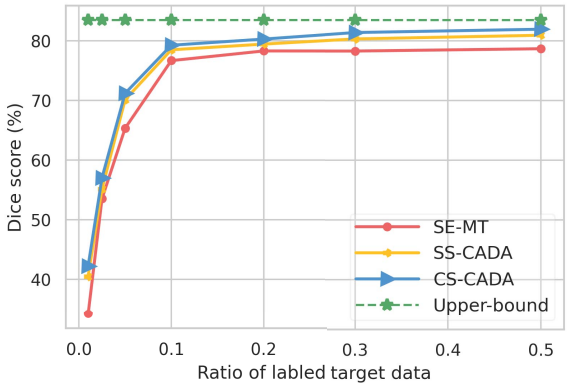

Fig. 4. Average Dice under different ratios of annotated XAs for coronary artery segmentation. The green dotted line is the upper bound of fully supervised learning.

of 2 patients to serve as T_L, and we took 121 XAs of 19 patients without annotations as T_U. The other 20 XAs of 3 patients and 36 XAs of 6 patients were used for validation and testing, respectively. For FIs, we only used the green channel, and all FIs and XAs were resized to 512 × 512. We used random horizontal and vertical flipping, random cropping and resize with an output size of 400 × 400 for data augmentation. The image intensity was normalized to [0, 1]. The batch size was 16, where 8, 4 and 4 images were from S_L, T_L and T_U, respectively. The first row of Fig. 1 shows a comparison of samples from two domains and their corresponding histograms.

2) Comparison With Existing Methods: To demonstrate the effectiveness of CS-CADA for solving cross-anatomy domain shift, we compared it with the baseline in two settings: 1) Using only S_L: A standard U-Net only learns from the source domain images, which is denoted as Baseline (source); 2) Using only T_L: A standard U-Net learns only from the labeled target domain images, which is denoted as Baseline (target). We also compared it with several state-of-the-art methods in four categories: 1) UDA methods that use S_L and unlabeled T_U: We investigated three UDA methods including ADDA [29], SIFA [6], and SC-GAN [11]; 2) Supervised Domain Adaptation (SDA) [7] methods that use S_L and T_L for training. We considered four methods, including Joint Training that takes $S_L \cup T_L$ as a single uniform training set, Fine-tuning [7] where the segmentation model is pre-trained on S_L and fine-tuned with T_L. Here, we consider two types of fine-tuning strategies: fine-tuning (last) means only updating parameters in last convolutional block of the decoder, and fine-tuning (all) means updating the whole set of parameters of the model. Both fine-tuning methods used 2000 iterations. X-shape [45] that separately learns from S_L and T_L, and DSBN [42] that uses domain-specific batch normalization for joint training; 3) SSL methods that use T_L and T_U for training, and we considered three state-of-the-art methods including SE-MT [46], UA-MT [21] and Cross Pseudo Supervision (CPS) [14]; and 4) Semi-Supervised Domain Adaptation (SSDA) method, i.e., Dual-teacher (Dual-T) [33]. All the compared methods were quantitatively evaluated with Recall, Precision and Dice score.

Comprehensive experimental results are shown in Table I. It presented that Baseline (source) only achieved an average Dice of 37.31%, showing the large domain gap between FIs and XAs. Baseline (target) only obtained average Dice at 69.81%, indicating that using a small set of labeled XAs cannot lead to accurate results. The UDA methods outperformed Baseline (source), and SC-GAN [11] was better than ADDA [29] and SIFA [6], but its performance is worse than baseline (target) due to the large domain gap and they do not use supervision from the target domain. For SDA methods, Fine-tuning (all) [7] achieved the highest Dice score of 75.50%. Joint Training, X-shape [45] and DSBN [42] based methods achieved better results than fine tuning the only last block of the decoder and UDA methods, demonstrating that small set of labeled XAs can provide effective supervision for bridging the cross-anatomy domain shift, in which the DSBN got the highest Dice (74.12%) among them. The SSL methods generally performed better than the SDA methods, showing the usefulness of unannotated images in the target domain. However, all these methods were inferior to our proposed CS-CADA that obtained an average Dice of 79.28%, which is a large improvement from 75.94% obtained by existing SSDA method Dual-T [33]. Our method also has higher Precision and Recall than the other methods as shown in Table I.

Fig. 3 is visual comparison between different methods in two cases. We selected SC-GAN [11], DSBN [42] and SE-MT [46] that are the best UDA, SDA and SSL method shown in Table I, respectively. SC-GAN exhibits obvious false negatives for thin terminal vessels, and false positives scattered in the background, which demonstrates that UDA methods designed for different domains with the same anatomical structure are not suitable for cross-anatomy adaptation. Compared with SC-GAN, DSBN obtained less false positives, but the connectivity of coronary arteries is not well preserved. SE-MT restored more details in the results, but the false negative region is visually obvious. In contrast, our CS-CADA obtained accurate and detailed results with very small amount of false negatives and false positives even for the thin terminals.

3) Ablations Study: For ablation study, we compared our CS-CADA with three variants: 1) SE-MT that does not use the source domain images therefore without DSBN and L_{ct}, 2) DSBN that only learns from S_L and T_L based on shared

TABLE III

QUANTITATIVE COMPARISON BETWEEN OUR CS-CADA AND STATE-OF-THE-ART METHODS FOR LV AND MYO SEGMENTATION. ONLY 10% OF TRAINING IMAGES IN THE TARGET DOMAIN ARE ANNOTATED. FINE-TUNING (LAST) MEANS ONLY UPDATING THE LAST CONVOLUTIONAL BLOCK OF THE DECODER. FINE-TUNING (ALL) MEANS UPDATING ALL THE PARAMETERS OF THE MODEL

Methods		Training set			LV		Myo		Average	
		S_L	T_U	T_L	Dice (%)	ASSD (mm)	Dice (%)	ASSD (mm)	Dice (%)	ASSD (mm)
Baseline (source)		✓			9.17±10.03	49.54±23.66	11.07±2.33	35.71±2.56	10.12±5.45	42.63±26.48
Baseline (target)				✓	77.87±6.63	6.96±3.35	54.56±7.93	8.37±2.98	66.21±4.66	7.67±3.67
UDA	ADDA [29]	✓	✓		40.59±23.30	8.57±9.51	20.32±11.60	12.98±4.85	30.45±16.38	10.78±7.94
	CycleGAN [35]	✓	✓		51.52±8.58	6.98±2.96	26.62±3.74	8.64±2.43	39.07±4.87	7.81±3.13
	SIFA [6]	✓	✓		78.81±11.14	5.97±2.74	41.13±7.08	9.44±1.94	59.97±7.72	7.71±2.95
SDA	Fine-tuning(last) [7]	✓		✓	45.43±11.07	25.22±6.80	26.06±9.54	20.60±6.09	35.75±9.68	22.91±6.85
	Fine-tuning(all)	✓		✓	85.51±5.37	8.79±6.02	71.74±4.89	6.00±3.48	78.63±4.44	7.40±5.11
	Joint Training	✓		✓	75.70±11.03	9.64±7.70	56.95±9.41	8.06±4.54	66.32±9.88	8.85±7.56
	X-shape [46]	✓		✓	82.92±9.38	3.39±2.35	67.91±6.44	4.44±2.30	75.41±7.39	3.92±3.27
	DSBN [43]	✓		✓	84.97±5.75	5.08±2.98	70.63±6.31	7.46±2.88	77.80±5.47	6.27±2.94
SSL	SE-MT [47]		✓	✓	82.10±5.39	5.29±3.21	69.18±4.84	5.46±2.63	75.65±4.66	5.38±3.27
	UA-MT [21]		✓	✓	81.56±6.13	5.40±3.74	71.95±5.49	5.40±1.51	76.76±3.74	4.40±3.03
	CPS [14]		✓	✓	82.53±10.42	4.18±3.25	67.39±7.61	4.70±2.44	74.96±8.54	4.44±2.89
SSDA	Dual-T [34]	✓	✓	✓	81.81±8.90	4.57±3.14	70.62±5.45	5.32±2.25	76.22±6.74	4.94±2.76
CS-CADA (ours)		✓	✓	✓	**87.02±4.77**	**3.15±2.20**	**74.85±5.90**	**3.37±2.04**	**80.80±4.94**	**3.26±2.15**

convolution with domain-specific batch normalization, and 3) a combination of SE-MT and DSBN, without using L_{ct}, which is denoted as SS-CADA [13]. The results in Table II show that SS-CADA combining S_L, T_L and T_U performed better than only learning from $T_L + T_U$ (SE-MT) and $S_L + T_L$ (DSBN). Finally, the comparison between SS-CADA and CS-CADA highlights the contribution of our proposed cross-domain contrastive learning strategy. The results show that by leveraging the existing data from another domain and the unannotated images in the target domain, our method improved the average Dice by 13.56% (i.e., from 69.81% to 79.28%). We conducted a statistical significance evaluation based on paired t-test between SS-CADA and CS-CADA. The p-value of Recall was $0.67 > 0.05$ meaning there is no significant difference. Although the p-value of Dice score was $0.076 > 0.05$, the p-value of Precision was $0.007 < 0.05$, showing the significant improvement from SS-CADA to CS-CADA. Visual comparison between SS-CADA and CS-CADA are shown in the last two columns of Fig. 3.

To investigate the annotation efficiency for our semi-supervised method with additional source domain images for training, we experimented with different ratios of labeled images in the target domain, i.e., 1%, 3%, 5%, 10%, 30% and 50% respectively. We compared our method with SE-MT and SS-CADA that are better than the other variants according to Table II. Fig. 4 shows that all these methods have a poor performance when the annotation ratio is below 10%, and the performance is much improved with a higher annotation ratio. Our proposed CS-CADA consistently outperformed SS-CADA and SE-MT with different annotation ratios. When the annotation ratio is 0.3 to 0.5, the performance of our CS-CADA is close to that of fully supervised learning, showing effectiveness of our method for reducing the annotation requirement in the target domain. Meanwhile, the performance gap between SE-MT and CS-CADA is larger when the annotation ratio is smaller, which reveals that our method have more advantages than existing methods when the amount of annotations are very limited (e.g., only less than 10% of the annotations are available).

C. Circular Structure Segmentation

1) Datasets: We then applied our method to Left Ventricle (LV) and left ventricular Myocardium (Myo) segmentation from MRI. As the LV and Myo are circular structures that have similar shapes with optic disc and cup, we use the REFUGE [47] dataset with annotated optic disc and cup as source domain images. It contains 400 retinal fundus images with a size of 2124×2056, acquired by a Zeiss Visucam 500 camera. As for glaucoma patients, the shapes of optic disc and cup may be abnormal, we only used the 360 non-glaucoma images in that dataset as our S_L. For preprocessing, we first cropped the image with a patch size of 640×640 centered on the optic disk and then resized it to 256×256. Meanwhile, we utilized the Multi-Sequence Cardiac MR segmentation challenge (MS-CMRSeg) dataset [48] as target domain images that consists of 45 multi-sequence CMR images from patients with cardiomyopathy. Each patient was scanned with LGE, T2-weighted and bSSFP sequences respectively, and we only used the bSSFP sequence for the segmentation task. The dataset has 412 2D slices in total, and we randomly split them into 282 slices of 32 patients for training, 36 slices of 4 patients for validation and 94 slices of 9 patients for testing. For the 282 training slices, 28 slices were used as T_L and the others were used as T_U. For preprocessing, the image was firstly cropped with a patch size of 160×160 centered on the target region and then resized to 256×256. All dataset were normalized the intensity to [0, 1] in training. The batch size was 24 where 12, 6 and 6 images were from S_L, T_L and T_U, respectively. The second row of Fig. 1 shows two examples of images in the source and target domains and their histograms.

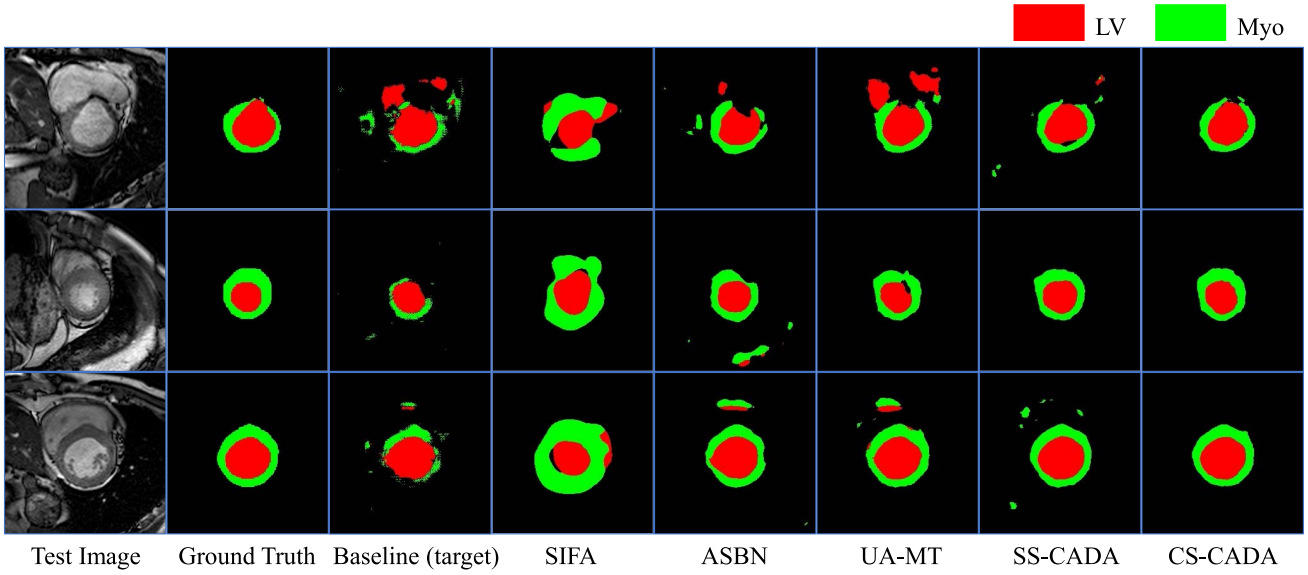

Fig. 5. Visual comparison of different methods for LV and Myo segmentation. Red and green regions denote the LV and Myo, respectively.

TABLE IV

QUANTITATIVE ABLATION STUDY OF OUR CS-CADA FOR LV AND MYO SEGMENTATION. SE-MT IS SEMI-SUPERVISED LEARNING ONLY USING T_L AND T_U. DSBN MEANS USING THE ANATOMY-SPECIFIC BATCH NORMALIZATION FOR LEARNING FROM S_L AND T_L. SS-CADA MEANS LEARNING FROM S_L, T_L AND T_U AT THE SAME TIME USING OUR PROPOSED DSBN AND SE-MT, WITHOUT USING CONTRASTIVE LEARNING

	LV		Myo		Average	
Methods	Dice (%)	ASSD (mm)	Dice (%)	ASSD (mm)	Dice (%)	ASSD (mm)
Baseline (target)	77.87±6.63	6.96±3.35	54.56±7.93	8.37±2.98	66.21±4.66	7.67±3.67
SE-MT	82.10±5.39	5.29±3.21	69.18±4.84	5.46±2.63	75.65±4.66	5.38±3.27
DSBN	84.97±5.75	5.08±2.98	70.63±6.31	7.46±2.88	77.80±5.47	6.27±2.94
SS-CADA [13]	86.76±5.88	3.37±2.17	71.74±6.88	4.30±2.19	79.25±5.94	3.84±2.21
CS-CADA (ours)	**87.02±4.77**	**3.15±2.20**	**74.85±5.90**	**3.37±2.04**	**80.80±4.94**	**3.26±2.15**

2) Comparison With Existing Methods: We employed almost the same set of methods as in Section IV-B.2 for comparison in the experiment. As the the SC-GAN was specifically designed for vessel segmentation, we replace it with CycleGAN [34] for comparison here. CycleGAN is a commonly used image-level alignment UDA method that utilizes cycle-consistent generative adversarial networks to achieve cross-modality synthesis. We employed Dice score and Average Symmetric Surface Distance (ASSD) for quantitative evaluation.

Quantitative evaluation results of the compared methods are shown in Table III. It can be observed that Baseline (source) obtained a very low average Dice of 10.12% for the LV and Myo segmentation, showing the large domain shift between retinal fundus images and CMR. Baseline (target) obtained an average Dice of 66.21%, which shows that only using the small number of annotated images in the target domain will limited the model's performance. For UDA methods, SIFA [6] performed better than ADDA [29] and CycleGAN [34], but its average Dice was only 59.97%, showing that classical UDA methods cannot be directly applied to cross-anatomy domain adaptation. For SDA methods, Fine-tuning (last) [7] had the worst performance with an average Dice of 35.75% while fine-tuning (all) achieved an average Dice score of 78.63%. This is very impressive and it shows that just fine-tuning is effective to achieve a better performance than learning directly from the small set of target images. However, its performance is lower than our CS-CADA that improved the average Dice to 80.80%. For the values for Joint Training, X-shape [45] and DSBN [42] were 66.32%, 75.41% and 77.80%, respectively. The results shows the effectiveness of DSBN to deal with the domain gap. The SSL methods are generally better than the UDA and SDA methods, and the average Dice for SE-MT [46], UA-MT [21] and CPS [14] were 75.65%, 76.76% and 74.96%, respectively. The excising SSDA method Dual-T [33] obtained an average Dice of 76.22%, which was similar to UA-MT and outperformed the other existing methods. In comparison, our proposed CS-CADA achieved higher performance than most above state-of-the-art methods, and it obtained an average Dice of 80.80% by combining DSBN and our proposed cross-domain contrastive learning in a semi-supervised framework. The ASSD values obtained by our CS-CADA for the LV and Myo were 3.15mm and 3.37mm, respectively, which were also superior in ASSD values of the other methods.

Fig. 5 shows a visual comparison between different methods for the LV and Myo segmentation, where we selected SIFA [6], DSBN [45] and UA-MT [21] that are the best UDA,

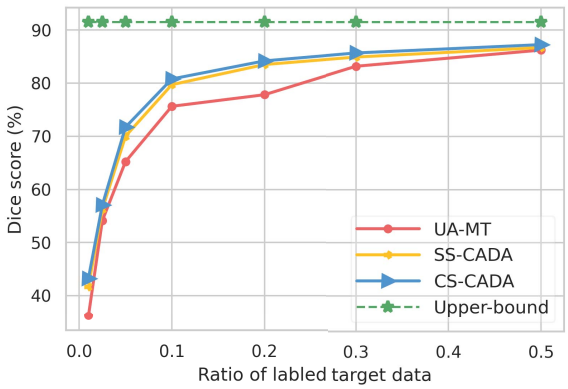

Fig. 6. Average Dice of LV and Myo achieved by different methods under different ratios of labeled images in the target domain. The green dotted line is the upper bound on fully supervised learning.

SDA and SSL methods according to Table III, respectively. It indicates that Baseline (target) achieved very poor results, with over-segmentation of LV and some missing part of Myo. SIFA captured the shape of LV and Myo through style transferring, but the result does not match the semantic boundary well and the Myo is noticeably over-segmented. In contrast, DSBN and UA-MT can better predict the overall structure than these two methods, but they gain a lot of false positive predictions. Our CS-CADA achieved much better results than the others, without false positives in the background, and the segmentation boundary is very close to that of the ground truth.

3) Ablations Study: In parallel with Section IV-B.3, we conducted the ablation study by comparing our CS-CADA with SE-MT that only leans from T_L and T_U, DSBN that only leans from S_L and T_L, and SS-CADA [13] that combines SE-MT and DSBN but does not use L_{ct}. The evaluation results of these methods are shown in Table IV. Compared with Baseline (target), SE-MT and DSBN improved the average Dice from 66.21% to 75.65% and 77.80%, respectively, which validates the effectiveness of them to leverage T_U and S_L, respectively. SS-CADA combining them together further improved the score to 79.25%. Finally, the better performance of CS-CADA than SS-CADA shows the effectiveness of our contrastive learning strategy. We also conducted statistical significance evaluation based on paired t-test between SS-CADA and CS-CADA. The p-value for the Dice score was $0.043 < 0.05$, showing the significant improvement from SS-CADA to CS-CADA. Note that with only 10% of training images in the target domain being annotated, our CS-CADA improved the average Dice score by 22% (from 66.21% to 80.80%).

We also compared our CS-CADA with UA-MT and SS-CADA under different ratios of annotated images in the target domain, and their average Dice scores are shown in Fig. 6. It can be observed that CS-CADA consistently outperformed UA-MT and SS-CADA. The gap between UA-MT and CS-CADA is larger when the annotation ratio is smaller, indicating that the effectiveness of our method leveraging additional dataset with similar structures in a different domain improves the segmentation performance when annotations in the target domain are limited.

V. Discussion and Conclusion

Reducing the annotation cost while maintaining the model's robust performance is important in medical image segmentation tasks because of the time-consuming and labor-intensive annotation process. Considering the availability of existing annotated datasets for a similar anatomical structure, our cross-anatomy domain adaptation method can improve the segmentation performance by transferring knowledge from such available datasets to the target segmentation task where only a small set of annotations are available for the target, which outperformed fine-tuning for transfer learning. Our method is also superior to existing semi-supervised methods that do not consider images from other domains, and better than existing domain adaptation methods that focus on the same anatomical structure in similar or different modalities. Note that in standard transfer learning setting, the pre-trained model is fine-tuned with the target dataset, without using the source dataset any more. If the source and target dataset have a large gap, the fine-tuning is less effective. In the DA setting of this work, the shared shape features can improve the performance on the target dataset effectively. Compared with pre-training with ImageNet, using datasets with similar shapes is more data and annotation-efficient.

Differently from classical domain adaptation considering only cross-modality domain shift for segmenting the same set of structures [6], [31], we deal with a more difficult scenario where an existing dataset with similar anatomical structures is used to assist model training in the target domain. Here, the "similar anatomical structure" requirement means there is a shape similarity between the two domains, i.e., the shapes have similar topologies but may be different in scales. For example, vessels in different organs are similar in terms of vascular shapes, but they can have different diameters and orientations. In the second scenario of our work, the LV and Myo in cardiac MRI and optic cup and disk in fundus images are obviously two different domains, but both of them have circular structures in different scales.

Differently from disentanglement methods [49], [50], the proposed CS-CADA does not need to specifically design modules to extract domain-invariant content and domain-specific style representations, respectively. Our CS-CADA captures anatomical representations through the shared convolutional layers and normalizes each style distribution to a common distribution space by DSBN. The whole procedure is conducted on a unified network. In contrast, disentanglement methods typically need to disentangle out the content and style representations through different networks and additionally introduce generative adversarial networks to discriminate them. These procedures are more complex and difficult to train compared with CS-CADA.

The term "across similar anatomical structures" is introduced for the aim of improving the model's performance on a target domain with the help of some easily available datasets with similar structures. Indeed, requiring "similar structures"

is more relaxed than the requirement of the exact set of objects in different domains (like CT and MRI) in typical DA setting. Though our method can also be applied to the same set of objects in different domains, this work makes it possible to adapt a model to a similar structure, rather than the same structure, which increases the chance of leveraging broader source domains. Compared with cross-domain contrastive learning, KL loss can also be used to encourage the distributions of $g_i^{\mathcal{S}}$ and $g_j^{\mathcal{T}}$ to be similar, but at the same time, we aim to encourage $(g_i^{\mathcal{S}}, g_i^{\mathcal{T}})$ and $(g_i^{\mathcal{S}}, g_j^{\mathcal{S}})$ to be divergent so that the BN layers can distinguish the different styles of two domains. Hence, using contrastive learning is more suitable for this purpose.

One limitation of this work is that it requires some annotated samples in the target domain for model training. As shown in Fig. 4 and Fig. 6, our proposed CS-CADA performed well when the annotation ratio changes from 30% to 50%, but in extreme situations (e.g., the annotation ratio is less than 10%), its performance drops dramatically. It may be a potential way to combine our method with a small-sample learning strategy (e.g., few/one-shot learning, test-time adaptation, etc.) to improve model adaptation ability in such situations. In addition, this work only deals with 2D images, and it is of interest to deal with 3D structures in the future.

In conclusion, we propose a Contrastive Semi-supervised learning for Cross Anatomy Domain Adaptation (CS-CADA) in medical image segmentation with knowledge transferred from an existing dataset to a target segmentation task, where they are different organs with similar anatomical structures. We use Domain-Specific Batch Normalization (DSBN) and shared convolutional kernels to jointly learn from the source and target domains, and propose cross-domain contrastive learning to learn domain-invariant features, which is combined with a self-ensembling mean teacher framework to further leverage unannotated images in the target domain. Experimental results show that our method can effectively achieve accurate segmentation of coronary artery from XAs and left ventricle and myocardium from MRI with limited annotations, which has a potential to reduce the annotation cost. In the future, it is of interest to apply our method to other segmentation tasks, and improve the robust performance under the extremely limited annotations.

ACKNOWLEDGMENT

The work was done during Ran Gu internship at SenseTime Research.

REFERENCES

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

[2] R. Gu *et al.*, "CA-Net: Comprehensive attention convolutional neural networks for explainable medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 40, no. 2, pp. 699–711, Nov. 2020.

[3] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. N. Chiang, Z. Wu, and X. Ding, "Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation," *Med. Image Anal.*, vol. 63, Jul. 2020, Art. no. 101693.

[4] M. Sajjadi, M. Javanmardi, and T. Tasdizen, "Regularization with stochastic transformations and perturbations for deep semi-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 1163–1171.

[5] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1195–1204.

[6] C. Chen, Q. Dou, H. Chen, J. Qin, and P.-A. Heng, "Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation," in *Proc. 33rd AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 865–872.

[7] N. Tajbakhsh *et al.*, "Convolutional neural networks for medical image analysis: Full training or fine tuning?" *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1299–1312, May 2016.

[8] Q. Dou, C. Ouyang, C. Chen, H. Chen, and P.-A. Heng, "Unsupervised cross-modality domain adaptation of ConvNets for biomedical image segmentations with adversarial loss," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 691–697.

[9] J. Staal, M. D. Abramoff, M. Niemeijer, M. A. Viergever, and B. van Ginneken, "Ridge-based vessel segmentation in color images of the retina," *IEEE Trans. Med. Imag.*, vol. 23, no. 4, pp. 501–509, Apr. 2004.

[10] A. Hoover and M. Goldbaum, "Locating the optic nerve in a retinal image using the fuzzy convergence of the blood vessels," *IEEE Trans. Med. Imag.*, vol. 22, no. 8, pp. 951–958, Aug. 2003.

[11] F. Yu *et al.*, "Annotation-free cardiac vessel segmentation via knowledge transfer from retinal images," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Cham, Switzerland: Springer, 2019, pp. 714–722.

[12] Y. Ganin *et al.*, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2030–2096, 2017.

[13] J. Zhang, R. Gu, G. Wang, H. Xie, and L. Gu, "SS-CADA: A semi-supervised cross-anatomy domain adaptation for coronary artery segmentation," in *Proc. IEEE 18th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2021, pp. 1227–1231.

[14] X. Chen, Y. Yuan, G. Zeng, and J. Wang, "Semi-supervised semantic segmentation with cross pseudo supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2613–2622.

[15] Y. Zhang, L. Yang, J. Chen, M. Fredericksen, D. P. Hughes, and D. Z. Chen, "Deep adversarial networks for biomedical image segmentation utilizing unannotated images," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Cham, Switzerland: Springer, 2017, pp. 408–416.

[16] D. Fan *et al.*, "Inf-Net: Automatic COVID-19 lung infection segmentation from CT images," *IEEE Trans. Med. Imag.*, vol. 39, no. 8, pp. 2626–2637, Aug. 2020.

[17] G. Wang *et al.*, "Semi-supervised segmentation of radiation-induced pulmonary fibrosis from lung CT scans with multi-scale guided dense attention," *IEEE Trans. Med. Imag.*, vol. 41, no. 3, pp. 531–542, Mar. 2022.

[18] H. Zheng *et al.*, "Semi-supervised segmentation of liver using adversarial learning with deep atlas prior," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Cham, Switzerland: Springer, 2019, pp. 148–156.

[19] G. Bortsova, F. Dubost, L. Hogeweg, I. Katramados, and M. de Bruijne, "Semi-supervised medical image segmentation via learning consistency under transformations," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Cham, Switzerland: Springer, 2019, pp. 810–818.

[20] X. Luo, J. Chen, T. Song, and G. Wang, "Semi-supervised medical image segmentation through dual-task consistency," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, 2021, pp. 8801–8809.

[21] L. Yu, S. Wang, X. Li, C.-W. Fu, and P.-A. Heng, "Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Cham, Switzerland: Springer, 2019, pp. 605–613.

[22] W. Cui *et al.*, "Semi-supervised brain lesion segmentation with an adapted mean teacher model," in *Proc. Int. Conf. Inf. Process. Med. Imag.* Cham, Switzerland: Springer, 2019, pp. 554–565.

[23] X. Li, L. Yu, H. Chen, C. Fu, and P. Heng, "Transformation-consistent self-ensembling model for semisupervised medical image segmentation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 2, pp. 523–534, Feb. 2021.

[24] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[25] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 3320–3328.

[26] N. Tajbakhsh *et al.*, "Convolutional neural networks for medical image analysis: Full training or fine tuning?" *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1299–1312, May 2016.

[27] H. Chen *et al.*, "Standard plane localization in fetal ultrasound via domain transferred deep neural networks," *IEEE J. Biomed. Health Inform.*, vol. 19, no. 5, pp. 1627–1636, Sep. 2015.

[28] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1180–1189.

[29] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7167–7176.

[30] S. Valverde *et al.*, "One-shot domain adaptation in multiple sclerosis lesion segmentation using convolutional neural networks," *NeuroImage, Clin.*, vol. 21, Jan. 2019, Art. no. 101638.

[31] J. Chen *et al.*, "Adaptive hierarchical dual consistency for semi-supervised left atrium segmentation on cross-domain data," *IEEE Trans. Med. Imag.*, vol. 41, no. 2, pp. 420–433, Feb. 2022.

[32] R. Dorent *et al.*, "Scribble-based domain adaptation via co-segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Cham, Switzerland: Springer, 2020, pp. 479–489.

[33] K. Li, S. Wang, L. Yu, and P.-A. Heng, "Dual-teacher: Integrating intra-domain and inter-domain teachers for annotation-efficient cardiac segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Cham, Switzerland: Springer, 2020, pp. 418–427.

[34] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2223–2232.

[35] K. Kamnitsas *et al.*, "Unsupervised domain adaptation in brain lesion segmentation with adversarial networks," in *Int. Conf. Inf. Process. Med. Imag.*, pp. 597–609, Springer, 2017.

[36] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jun. 2006, pp. 1735–1742.

[37] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[38] K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu, "Contrastive learning of global and local features for medical image segmentation with limited annotations," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 12546–12558.

[39] G. Kang, L. Jiang, Y. Yang, and A. G. Hauptmann, "Contrastive adaptation network for unsupervised domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4893–4902.

[40] A. Singh, "CLDA: Contrastive learning for semi-supervised domain adaptation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 5089–5101.

[41] 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*. Cham, Switzerland: Springer, 2015, pp. 234–241.

[42] W.-G. Chang, T. You, S. Seo, S. Kwak, and B. Han, "Domain-specific batch normalization for unsupervised domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7354–7362.

[43] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–13.

[44] X. Wang, R. Zhang, C. Shen, T. Kong, and L. Li, "Dense contrastive learning for self-supervised visual pre-training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3024–3033.

[45] V. V. Valindria *et al.*, "Multi-modal learning from unpaired images: Application to multi-organ segmentation in CT and MRI," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 547–556.

[46] G. French, M. Mackiewicz, and M. Fisher, "Self-ensembling for visual domain adaptation," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–15.

[47] J. I. Orlando *et al.*, "REFUGE challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs," *Med. Image Anal.*, vol. 59, Jan. 2020, Art. no. 101570.

[48] X. Zhuang, "Multivariate mixture model for myocardial segmentation combining multi-source images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 12, pp. 2933–2946, Dec. 2019.

[49] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, "Diverse image-to-image translation via disentangled representations," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 35–51.

[50] S. Benaim, M. Khaitov, T. Galanti, and L. Wolf, "Domain intersection and domain difference," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3445–3453.