Unsupervised Domain Adaptation Network With Category-Centric Prototype Aligner for Biomedical Image Segmentation

PING GONG[1], WENWEN YU[1], QIUWEN SUN[1], RUOHAN ZHAO[2], AND JUNFENG HU[1]

[1]School of Medical Imaging, Xuzhou Medical University, Xuzhou 221000, China
[2]Department of Computing, The Hong Kong Polytechnic University, Hong Kong

Corresponding author: Junfeng Hu (hjf098@xzhmu.edu.cn)

ABSTRACT With the widespread success of deep learning in biomedical image segmentation, domain shift becomes a critical and challenging problem, as the gap between two domains can severely affect model performance when deployed to unseen data with heterogeneous features. To alleviate this problem, we present a novel unsupervised domain adaptation network, for generalizing models learned from the labeled source domain to the unlabeled target domain for cross-modality biomedical image segmentation. Specifically, our approach consists of two key modules, a conditional domain discriminator (CDD) and a category-centric prototype aligner (CCPA). The CDD, extended from conditional domain adversarial networks in classifier tasks, is effective and robust in handling complex cross-modality biomedical images. The CCPA, improved from the graph-induced prototype alignment mechanism in cross-domain object detection, can exploit precise instance-level features through an elaborate prototype representation. In addition, it can address the negative effect of class imbalance via entropy-based loss. Extensive experiments on a public benchmark for the cardiac substructure segmentation task demonstrate that our method significantly improves performance on the target domain.

INDEX TERMS Biomedical image segmentation, cross-modality learning, unsupervised domain adaptation, category-centric prototype aligner.

I. INTRODUCTION

Deep neural networks have achieved remarkable success in recent years; a variety of challenging medical imaging problems, such as biomedical image segmentation [1], [2], have witnessed breakthroughs, when a large quantity of labeled data are used, and the training and testing data are sampled from the sample distribution [3]. However, deep neural networks deployed in real-life applications usually suffer from the domain shift problem [4]. In biomedical imaging scenarios, this problem is even more obvious, as biomedical images have very different characteristics when they are acquired with different acquisition parameters or modalities [5], [6], such as magnetic resonance imaging (MRI) and computed tomography (CT). In addition, manual annotation in the field

of biomedical images is a time-consuming and expensive task. Compared to natural images, domain adaptation is more challenging on cross-modality biomedical images.

To address these problems, unsupervised domain adaptation (UDA) has been intensively studied to generalize the well-trained models on unlabeled target data, which seeks to transfer knowledge from labeled training data (source domain) to test data (target domain). To obtain domain-invariant feature representations across both domains, common methods for domain adaptation can be roughly divided into two types: 1) measures for minimizing an explicitly defined distance [7], [8], and 2) applying adversarial training to align the latent feature space [9], [10].

For instance, Gretton *et al.* [11] minimized the maximum mean discrepancy (MMD) distance between the source and target domains. Long *et al.* [8] proposed the multikernel MMD distance. Other works based on adversarial training
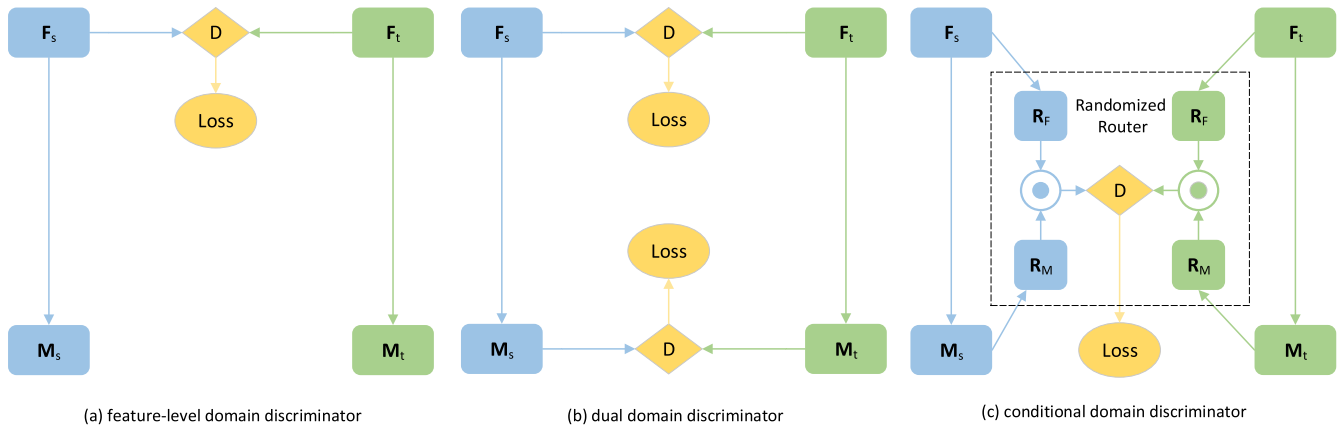
FIGURE 1. Frameworks of different domain discriminator. (a) feature-level domain discriminator based method. (b) dual-domain discriminator (feature-level and instance-level)-based method. (c) our proposed models named the conditional domain discriminator.

employed a domain classifier to facilitate domain invariance on the input level [12], feature level [9] (Figure 1(a)), output level and their combinations [6] (Figure 1(b)). Nevertheless, how to balance the ratio between multilevel discriminators requires empirical design and time for UDA and can severely affect the domain adaptation performance. Furthermore, multilevel discriminators may come at a contradiction in the adversarial adaptation procedure given the complex combinations of cross-modality between domains. In addition, domain classifiers tend to capture global-level discrepancy features across both domains, ignoring distinct modal information of instances in the cross-modality scenario, especially when neighboring structures contain unclear boundaries or relatively homogeneous tissues [6], and combining the feature-level and instance-level alignment of the source and target domains makes it difficult to perfectly address domain shift [13]–[16].

Motivated by these problems, we propose the unsupervised domain adaptation network with category-centric prototype aligner for biomedical image segmentation illustrated in Figure 2. Specifically, we introduce two key components, a conditional domain discriminator (CDD) and a category-centric prototype aligner (CCPA). Our method incorporates the CDD module inspired by [17] into the existing architecture to model the multimodal information and joint distributions of multilevel features via a randomized router, which makes it easier to fully capture multiplicative interactions between feature representation and segmentation prediction for domain adaptation. In CCPA, we extend category-level prototype alignment to cross-domain segmentation tasks, inspired by [13], [14], to guarantee the discriminability by exploiting more precise instance-level features. In addition, it can mitigate the negative effect of class imbalance on domain adaptation via an entropy-based loss to control the process of adaptation during the training phase.

The main contributions of this paper can be summarized as follows:

- In this paper, we present a novel framework for UDA, which is more effective and robust in handling complex cross-modality biomedical image segmentation. It efficiently balances multi-level features discriminator without contradiction that is crucial for UDA.
- We introduce CCPA module into the image segmentation framework, which exploits more precise instance-level features. In addition, it also can mitigate the negative effect of class-imbalance via entropy-based loss.
- We conduct experiments on one public benchmark for cardiac substructure segmentation tasks and show that our method significantly improves performance on the target domain.

II. RELATED WORK
The purpose of UDA is to generalize the model learned from the labeled source domain to the unlabeled target domain. A large number of adaptive methods have been proposed from different perspectives, including input-level, feature-level, output-level adaptation, and their combinations. In academia, UDA can be divided into two categories: 1) measures for minimizing a specific domain discrepancy metric, and 2) applying adversarial training to align the latent features space. In this section, we will give a brief review of related works in both areas. A more detailed review of UDA for image segmentation can be found in [18]–[22].

In the field of UDA, early research mainly focused on aligning the distributions of feature space by minimizing distance measurements. For example, Tzeng *et al.* [23] used the MMD distance as a minimized target between the source and target domains. Shen *et al.* [24] learned domain-invariant feature representations via Wasserstein distance guided representation learning. Long *et al.* [8] presented the multikernel MMD distance based on MMD to reduce the domain discrepancy. Yan *et al.* [7] further extended the work and employed weighted MMD with a task-specific loss for domain adaptation. Ding *et al.* [25] proposed an adaptive exploration method to maximize the distances of all target images, minimizes distances of similar target images, and address the domain-shift problem for
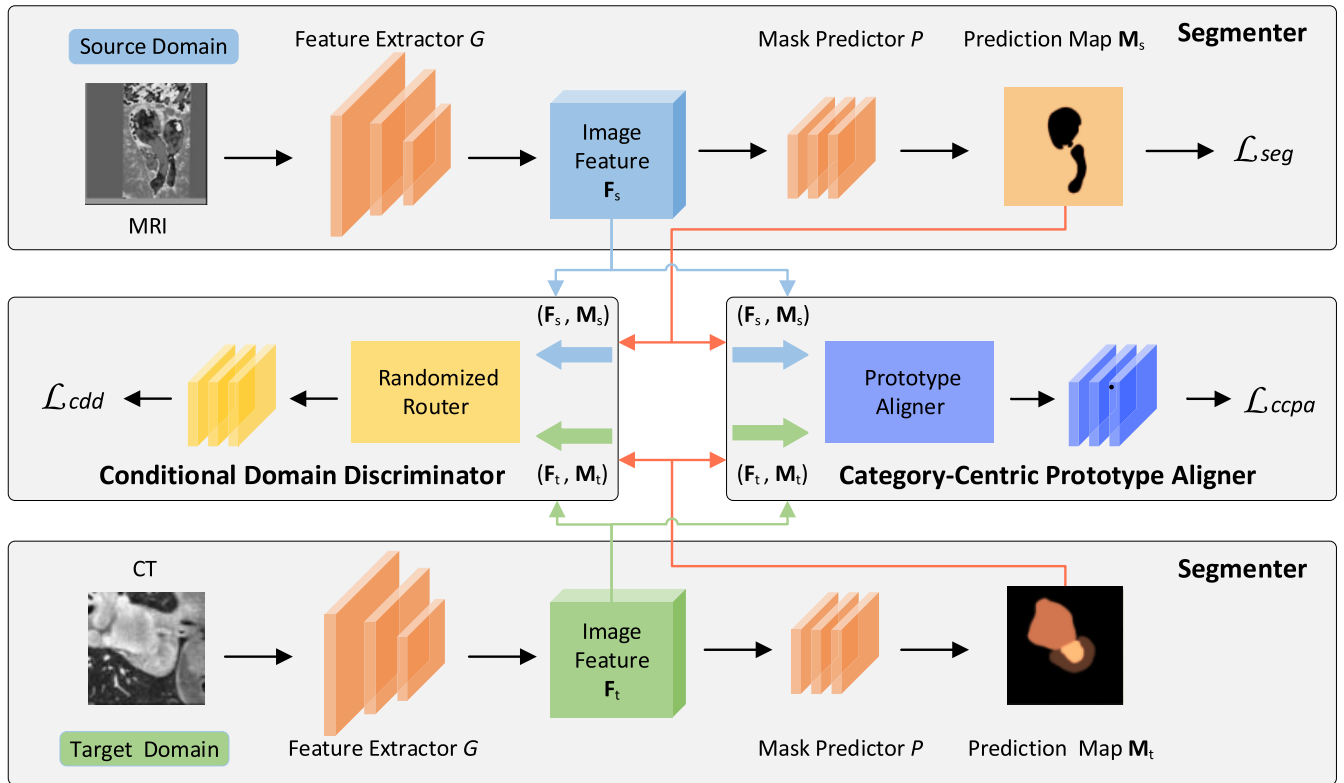
FIGURE 2. Overview of the framework. The segmenter is shared for both the source and target domains. The conditional domain discriminator and category-centric prototype aligner are trained from both source and unlabeled target data. During inference, the target domain is predicted by the segmenter.

person re-identification (re-ID) in an unsupervised manner. Fan *et al.* [26] proposed a progressive unsupervised learning method to transfer pretrained deep representations to unseen domains based on clustering to improve the re-ID accuracy and produced CNN models with high discriminative ability. Nevertheless, both [25], [26] are aimed at the topic of re-ID. In addition, the difference between [25], [26] and our method is that [25] minimizes distances between one image and its neighbors and maximizes distances between one image and other images at the feature level, and [26] performs clustering at the feature level. However, our method is not only performs conditional domain discriminator at the feature level and instance level, but also aligns category-centric prototype at the instance-level.

More recently, with the advent of generative adversarial networks (GAN) [27], another line of research is based on adversarial training. For instance, Ganin and Lempitsky [9] aligned the distributions of features across the two domains accomplished through standard backpropagation training via a simple new gradient reversal layer. Tzeng *et al.* [28] introduced a more flexible adversarial learning framework with united weight sharing to address the problem of domain shift. With the widespread success of CycleGAN [29] in unpaired image-to-image transformations, many previous image adaptation efforts were based on modified CycleGAN with applications in both natural datasets [30], [31] and

medical image segmentation [32]–[34]. Overall, our framework belongs to the latter category based on adversarial learning.

For biomedical image segmentation applications, as the common cross-modality, interscanner, and different imaging protocols vary, domain shift has become a critical and challenging problem in biomedical image segmentation. Ghafoorian *et al.* [35] reduced the required number of labels in the target domain via transfer learning for brain MRI lesion segmentation tasks. Opbroek *et al.* [36] proposed four transfer classifiers to deal with differences in distributions between training and target data and improved performance over supervised learning for segmentation across scanners and scan protocols. Nevertheless, these methods require extra annotations of target data. Instead, unsupervised domain adaptation is more desirable due to the requirements of zero additional target domain labels. Recent literature shows the great potential of unsupervised domain adaptation for biomedical image segmentation. For example, Kamnitsas *et al.* [37] made the earliest attempts to align feature distributions with an adversarial loss for unsupervised domain adaptation cross-protocol MRI segmentation and achieved promising adaptation performance. Degel *et al.* [38] and Zhang *et al.* [39] combined regularization with adversarial training and obtain better adaptation results on ultrasound datasets and cardiac MRI segmentation, respectively.

Wang *et al*. [40] presented a novel patch-based output space adversarial learning framework to jointly and robustly segment the optic disc and optic cup from different fundus image datasets and achieved effective feature alignment. However, these works did not aim at the topic of unsupervised domain adaptation for cross-modality biomedical images. In the field of cross-modality biomedical segmentation, Dou *et al*. [5], [6] employed adversarial learning to adapt the early-layer feature distributions while the higher-layer features are fixed. However, their method needs comprehensive empirical studies and design to determine the optimal adaptation depth. Chen *et al*. [33], [34] used deeply synergistic image and feature alignment for unsupervised bidirectional cross-modality adaptation segmentation and achieved better domain performance, but it has a complex training process due to the essence of GAN is used.

The most related works to our method are [13], [17] based on the graph-inducted prototype alignment mechanism and conditional domain adversarial networks, but still differ from our method in several aspects. These works are not used for the field of unsupervised domain adaptation for cross-modal biomedical image segmentation. More specifically, the difference between [17] and our proposed CDD module is that conditional adversarial domain adaptation is a common operation in the classifier tasks but rarely in segmentation tasks, especially in the field of unsupervised domain adaptation on cross-modal medical image segmentation. The difference between [13] and the proposed CCPA is that [13] is only suitable for object detection tasks because it performs graph-induced prototype alignment on the bounding-box level. However, the proposed CCPA module extends [13] to the pixel-level for segmentation tasks. We use a learnable graph attention layer to obtain the adjacency matrix of the graph at the pixel-level instead of calculating weights between region proposals at the bounding-box level.

III. METHOD
In this section, we provide a detailed description of our proposed method for UDA in biomedical image segmentation. Figure 2 gives the overall architecture of our method, which contains 3 modules:

- *Segmenter*: This module consists of the feature extractor G, and the mask predictor P. The feature extractor G encodes source/target domain (MRI/CT) images using CNN to obtain image features $\mathbf{F}_s/\mathbf{F}_t$. $\mathbf{F}_s$ and $\mathbf{F}_t$ represent semantic information of the source domain and target domain in the latent feature space, respectively. Then the mask predictor P uses image features $\mathbf{F}_s/\mathbf{F}_t$ as input to predict the segmentation map $\mathbf{M}_s/\mathbf{M}_t$ of the source domain and target domain individually.
- *Conditional Domain Discriminator*: This module can capture the latent relation between image feature $\mathbf{F}$ and prediction map $\mathbf{M}$ through a randomized router, in which image features interact with the prediction map. In addition, it can capture the cross-covariance between

image feature $\mathbf{F}$ and prediction map $\mathbf{M}$ to improve the transferability.
- *Category-Centric Prototype Aligner*: This module performs category-level prototype alignment to guarantee discriminability by exploiting more precise instance-level features. In addition, it can mitigate the negative effect of class imbalance on domain adaptation via an entropy-based loss to control the process of adaptation during the training phase. Thus, CCPA can obtain the capacity of transfer from the source domain to the target domain through UDA.

To ease understanding, our full model is described in parts. First, we begin by introducing the notation used in this paper in Section III-A. Our segmenter representation is described in Section III-B; then, the proposed CDD module mechanism is described in Section III-C. Finally, Section III-D shows how the CCPA module works.

A. NOTATION
In UDA, given N^s labeled examples of the source domain and N^t unlabeled examples of the target domain, they are denoted by $\mathcal{D}_s = \left\{ (x_1^s, y_1^s), \ldots, (x_{N^s}^s, y_{N^s}^s) \right\}$ and $\mathcal{D}_t = \left\{ x_1^t, \ldots, x_{N^t}^t \right\}$, respectively, where x_i^s is the image for the *i*-th sample of the source domain and follows distributions $\mathbb{P}_s$, and y_i^s represents its corresponding segmentation one-hot label. Similarly, x_i^t follows distributions $\mathbb{P}_t$, and the i.i.d. assumption is violated as $\mathbb{P}_s \neq \mathbb{P}_t$.

For ease of notation, we directly use x^s and y^s to represent the sample and label from the source domain, where we omit the subscript index *i* in the following subsections.

B. SEGMENTER
As shown in Figure 2, the top position in the diagram is the segmenter module, which contains two steps for the source domain. Specifically, image sample x^s is first forwarded into the feature extractor G, yielding image features $\mathbf{F}_s \in \mathbb{R}^{H \times W \times d}$ in latent feature space, where H, W, d denotes the height and width of the output map, and the dimension of feature embedding, respectively. Next, to perform segmentation on image features, the mask predictor P learns to create a mapping from image features to the label space $\mathcal{Y}^s$ by conducting supervised learning. Formally, prediction map $\mathbf{M}_s \in \mathbb{R}^{H \times W \times N_c}$, where N_c denotes the number of classes, is defined as follows :

$$\begin{cases} \mathbf{M}_s = P\left(\mathbf{F}_s ; \Theta_{\mathrm{p}}\right), \\ \mathbf{F}_s = G\left(x^s ; \Theta_{\mathrm{g}}\right), \end{cases} \tag{1}$$

where Θ_{p} and Θ_{g} represent the learned parameters of the mask predictor and the feature extractor, respectively. Similarly, as shown in the bottom position of Figure 2, the target domain sample x^t also has the same procedure to generate $\mathbf{F}_t \in \mathbb{R}^{H \times W \times d}$ and $\mathbf{M}_t \in \mathbb{R}^{H \times W \times N_c}$ during the domain adaptation phase, which is also fed into CDD and CCPA. In addition, Θ_{p} and Θ_{g} are shared parameters for both the source and target domains. In practice, we apply dilated

residual blocks [41] to the feature extractor G to obtain a large receptive field and preserve the spatial acuity of feature maps. For dense predictions in the segmentation task, the upsampling operation following by a softmax layer is used in the mask predictor P for probability predictions of the pixels.

To transfer the label information from the source domain to the target domain, supervised learning with the source domain $\mathcal{D}_s$ is an essential part of domain adaptation. The $\mathcal{L}_{seg}$ loss is defined as follows

$$\mathcal{L}_{seg} = -\frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{c=1}^{N_c} y_s^{(h,w,c)} \log(M_s^{(h,w,c)})$$
$$- \eta \sum_{c=1}^{N_c} \frac{\sum_{h=1}^{H}\sum_{w=1}^{W} 2 y_s^{(h,w,c)} M_s^{(h,w,c)}}{\sum_{h=1}^{H}\sum_{w=1}^{W} y_s^{(h,w,c)} + \sum_{h=1}^{H}\sum_{w=1}^{W} M_s^{(h,w,c)}}, \quad (2)$$

where $\mathbf{y}_s \in \mathbb{R}^{H \times W \times N_c}$ denotes one-hot labels, and $\eta \in [0, 1]$ is a trade-off parameter. The first term is the cross-entropy loss for pixelwise classification. The second term is the Dice loss for multiple cardiac structures, which is commonly employed in biomedical image segmentation tasks. We combine the two complementary hybrid loss functions to address the challenging heart segmentation task.

C. CONDITIONAL DOMAIN DISCRIMINATOR

Existing UDA works [5], [6] use dual-domain discriminators whose inputs are the feature-level features and the predicted segmentation masks of the source and target domains respectively. These works aim to align the latent feature space of the target domain to that of the source domain, with a more explicit constraint on the shape of segmentation masks. Although the starting point for these works [5], [6] is good and performs well on datasets, how to balance the ratio between dual discriminators can severely affect the domain adaptation performance, as it needs empirical design and time for UDA. Additionally, due to the lack of the supervisory signal of the target domain, the predicted map $\mathbf{M}_t$ probably have error noise or inaccurate boundaries, which makes it harder to align the source domain and target domain on the instance level when data distributions embody complex multimodal structures.

In this regard, we incorporate CDD inspired by [17] into the existing architecture illustrated in the left-middle position of Figure 2 to effectively align different domains of multimodal distributions native in segmentation problems. This is different from aligning the features and segmentation maps separately [42], [43]. Notably, this module can capture the cross-covariance between feature representations and segmentation predictions to improve the transferability. The key to CDD models is a novel conditional domain discriminator conditioned on the cross-covariance of domain-specific feature representations and segmentation predictions.

Figure 1(c) shows a detailed description of the CDD. Given two inputs $\mathbf{F}$ and $\mathbf{M}$ of the source domain and target domain

respectively, the CDD models the multimodal information and joint distributions of $\mathbf{F}$ and $\mathbf{M}$ via a randomized router, which makes it easier to fully capture multiplicative interactions between feature representation and segmentation prediction. Then, conditional domain adversarial loss measures the domain discrepancy by training a domain discriminator in a conditional manner. Given conditional domain discriminator D_{cdd}, the definition of conditional domain adversarial loss is

$$\mathcal{L}_{cdd}(\mathcal{D}_s, \mathcal{D}_t) = \mathbb{E}_{x^s \sim \mathcal{D}_s} \log[D_{cdd}(J(\mathbf{F}_s, \mathbf{M}_s))]$$
$$+ \mathbb{E}_{x^t \sim \mathcal{D}_t} \log[D_{cdd}(J(\mathbf{F}_t, \mathbf{M}_t))], \quad (3)$$

where J is the explicit randomized multilinear map that converts $\mathbf{F}$ and $\mathbf{M}$ to a single tensor that is the joint variable of domain-specific feature representation and segmentation prediction for adversarial adaptation. We define

$$J(\mathbf{F}, \mathbf{M}) = \frac{1}{\sqrt{d_o}} (\mathbf{F}\mathbf{R_F}) \odot (\mathbf{M}\mathbf{R_M}) \in \mathbb{R}^{H \times W \times d_o}, \quad (4)$$

where $\odot$ represents an element-wise product, d_o is the dimension of the output tensor, and $\mathbf{R_F} \in \mathbb{R}^{d \times d_o}$ and $\mathbf{R_M} \in \mathbb{R}^{N_c \times d_o}$ are random matrices, in which each element R_{ij} follows a symmetric distribution, where $\mathbb{E}\left[R_{ij}\right] = 0$ and $\mathbb{E}[R_{ij}^2] = 1$. In practice, $\mathbf{R_F}$ and $\mathbf{R_M}$ are sampled only once from a uniform distribution or Gaussian distribution and fixed in the training phase. We restrict $d_o \ll d \times N_c$ of random matrices to avoid dimension explosion [17], [44], [45].

One common optimization approach for adversarial learning networks in an unsupervised manner is to follow the training rules of GAN [46], which is implemented in the form of a minimax two-player game containing a generator and a discriminator role. An alternative optimization method utilizes a gradient reversal layer (GRL) [9], which is inserted between the feature generator and CDD. In practice, we use the latter implementation method to align the domain statistics with CDD that predicts the domain.

D. CATEGORY-CENTRIC PROTOTYPE ALIGNER

The gap between the two domains can severely decrease the model's performance. Furthermore, as different instances commonly embody distinct modal information in the cross-modality scenario, especially when neighboring structures remain unclear boundaries or relatively homogeneous tissues [6], combining the feature-level and instance-level alignment of the source and target domains makes it difficult to perfectly address domain shift [13]–[16]. In this regard, in addition to the CDD module, we also proposed CCPA shown in the right-middle position of Figure 2, inspired by [13], which includes five steps to align the source and target domains with category-centric prototype representations. The detailed flowchart is illustrated in Figure 3. In addition, to relieve the negative effect of class imbalance [47] on domain adaptation, we design a category-reweighted contrastive loss to balance the training process of domain adaptation.
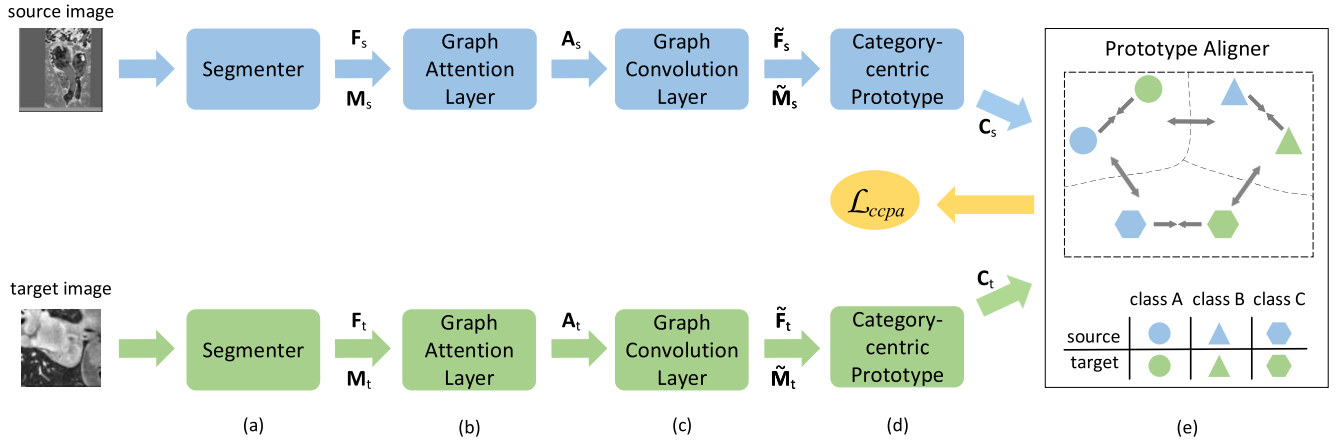
FIGURE 3. Category-Centric Prototype Aligner flowchart. (a) Feature representation **F** and segmentation prediction **M** are generated via the segmenter. (b) Pixel-level relational weights **A** are obtained by the graph attention layer. (c) More accurate instance-level feature representations $\widetilde{\mathbf{F}}$ and $\widetilde{\mathbf{M}}$ are acquired through the graph convolution layer. (d) Category-centric prototype **C** is derived via confidence-guided merging. (e) Performing category-level domain alignment by minimizing loss $\mathcal{L}_{ccpa}$ calculated by prototype aligner.

1) GRAPH ATTENTION LAYER

As described in Section III-B, feature representation **F** and segmentation prediction **M** are generated via a segmenter on both the source and target domains, which is the first step of the flowchart shown in Figure 3(a). Then, we use the graph attention layer [48], [49] to produce adjacency matrix $\mathbf{A} \in \mathbb{R}^{HW \times HW}$, which is used to model the relationship between pixels. Intuitively, two spatially closer pixels are more likely to depict the same class and should be assigned higher connection weights. Following this idea, a method for obtaining an adjacency matrix is defined as

$$\begin{cases} A_{ij} = \text{softmax}_j(\mathbf{e}_{ij}), & i, j \in [1, HW], \\ \mathbf{e}_{ij} = \text{LeakRelu}(\mathbf{w}_i^T |\mathbf{h}_i - \mathbf{h}_j|), \end{cases} \quad (5)$$

where $\mathbf{w}_i \in \mathbb{R}^d$ is the learnable weight vector, $\mathbf{h}_i \in \mathbb{R}^d$ and $\mathbf{F} = \{\mathbf{h}_i\}_1^{HW}$. To solve the problem of gradients vanishing in the training phase, we use LeakyReLU instead of the Relu activation function. In practice, to avoid high computing complexity and memory, we perform graph attention on downsample feature representation **F** and then upsample matrix **A** to the original resolution size.

2) GRAPH CONVOLUTION LAYER

Because of the boundary deviation in segmentation prediction, often distributed around the ground truth mask, initial segmentation prediction conveys incomplete instance information, which leads to inaccuracy of representing an instance. A natural approach here is that the segmentation prediction feature belonging to a certain instance should be aggregated, to achieve exact instance-level feature representations. Specifically, more exact instance-level feature representations are calculated by using an adjacency matrix **A** containing the spatial relevance, image features **F** and segmentation confidence **M**. Formulation is defined as

$$\widetilde{\mathbf{F}} = \sigma(\mathbf{A}\mathbf{F}\mathbf{W}_F) \in \mathbb{R}^{HW \times d}, \quad (6)$$
$$\widetilde{\mathbf{M}} = \sigma(\mathbf{A}\mathbf{M}\mathbf{W}_M) \in \mathbb{R}^{HW \times N_c}, \quad (7)$$

where $\mathbf{W}_F \in \mathbb{R}^{d \times d}$ and $\mathbf{W}_M \in \mathbb{R}^{N_c \times N_c}$ are learnable weight matrices. $\sigma(\cdot) = \max(0, \cdot)$ is a nonlinear activation function. In Eq. 6, 7, after graph convolution, $\widetilde{\mathbf{F}}$ and $\widetilde{\mathbf{M}}$ more precise instance-level information is aggregated through information propagation among adjacent pixels.

3) CATEGORY-CENTRIC PROTOTYPE

After obtaining more accurate feature representations aggregated on the instance level, we employ confidence-guided metering to integrate the multimodal information reflected by different instances into prototype representations. Category-centric prototypes $\mathbf{C} \in \mathbb{R}^{N_c \times d}$ are calculated by the weighted mean feature representation $\widetilde{\mathbf{F}}$:

$$\begin{cases} \mathbf{C} = \boldsymbol{\alpha}^T \widetilde{\mathbf{F}} \in \mathbb{R}^{N_c \times d}, \\ \alpha_{i,j} = \dfrac{\widetilde{P}_{i,j}}{\sum_{i=1}^{HW} \widetilde{P}_{i,j}}, \end{cases} \quad (8)$$

where $\boldsymbol{\alpha} \in \mathbb{R}^{HW \times N_c}$ represents normalized segmentation confidence to each class. The derived prototypes **C** serve as the proxy of each class during subsequent domain alignment.

4) PROTOTYPE ALIGNER

Following the heuristic rules of the prototype-based method [13], [15] for unsupervised domain alignment, we minimize intraclass loss to narrow the distance between the same categories' prototypes of two domains, named $\mathcal{L}_{intra}$. In addition, we also minimize interclass loss to bound the distance between different categories' prototypes, named $\mathcal{L}_{inter}$. Furthermore, in segmentation tasks, as a class-imbalance problem usually exists [47], we address this problem through reweight loss $\mathcal{L}_{intra}$ and $\mathcal{L}_{inter}$ using an entropy-based map [50]. The underlying idea is that hard samples or sample-scarce categories produce high-entropy predictions on both source and target domains because the categories with abundant samples are trained more sufficiently and better aligned so that they have higher

confidence compared with sample-scarce categories, and vice versa.

In this regard, we assign higher weights to the sample-scarce categories during the training process of domain adaptation. The weight corresponding to the k-th category is calculated by:

$$\beta_k = -\frac{1}{HW} \sum_{h=1}^{H} \sum_{w=1}^{W} M^{(h,w,k)} \log(M^{(h,w,k)}). \quad (9)$$

Note that the intraclass loss requires prototypes of the same category to be as close as possible, and the interclass loss constrains the distance between prototypes of different classes to be larger than a margin. The category-centric prototype domain adaptation loss $\mathcal{L}_{ccpa}$ consists of an intraclass loss and three interclass losses in which all pairwise relations between two domains' prototypes are considered. So we define the following loss:

$$\mathcal{L}_{ccpa} = \mathcal{L}_{intra}(\mathcal{D}_s, \mathcal{D}_t) + \frac{1}{3} \big(\mathcal{L}_{inter}(\mathcal{D}_s, \mathcal{D}_s)$$
$$+ \mathcal{L}_{inter}(\mathcal{D}_s, \mathcal{D}_t) + \mathcal{L}_{inter}(\mathcal{D}_t, \mathcal{D}_t) \big), \quad (10)$$

$$\mathcal{L}_{intra}(\mathcal{D}_s, \mathcal{D}_t) = \frac{\sum_{i=1}^{N_c} \beta_i^s \beta_i^t \Phi(C_i^s, C_i^t)}{\sum_{i=1}^{N_c} \beta_i^s \beta_i^t}, \quad (11)$$

$$\mathcal{L}_{inter}(\mathcal{D}, \mathcal{D}') = \frac{\sum_{1 \leqslant i \neq j \leqslant N_c} \beta_i^{\mathcal{D}} \beta_j^{\mathcal{D}'} max(0, m - \Phi(C_i^{\mathcal{D}}, C_j^{\mathcal{D}'}))}{\sum_{1 \leqslant i \neq j \leqslant N_c} \beta_i^{\mathcal{D}} \beta_j^{\mathcal{D}'}}, \quad (12)$$

where $\Phi(c, c') = ||c - c'||_2$ is the Euclidean distance between two prototypes and $\{C_i^s\}_{i=1}^{N_c}$, $\{C_i^t\}_{i=1}^{N_c}$ represent the prototypes of source and target domains. $\mathcal{D}$ and $\mathcal{D}'$ denote two domains from which prototype pairs belonging to different categories are drawn. m is the margin term which is fixed as 1.0 in all experiments.

Our model parameters of whole networks are jointly trained by minimizing the following total loss:

$$\mathcal{L}_{total} = \mathcal{L}_{seg} + \lambda_1 \mathcal{L}_{cdd} + \lambda_2 \mathcal{L}_{ccpa}, \quad (13)$$

where $\mathcal{L}_{seg}$, $\mathcal{L}_{cdd}$, and $\mathcal{L}_{ccpa}$ are defined in Eqs. 2, 3, and 10, respectively. λ_1 and $\lambda_2 \in [0, 1]$ are trade-off parameters.

IV. EXPERIMENTS
A. EXPERIMENTAL DETAILS
1) DATASETS
We use a medical cross-modality domain adaptation benchmark proposed in [6] to validate the performance of our proposed unsupervised cross-modality domain adaptation method for biomedical image segmentation. This dataset contains training (16 subjects) and testing (4 subjects) sets for each modality, and it collects from the public dataset of *MICCAI 2017 Multi-Modality Whole-Heart Segmentation* [51], which consists of 20 unpaired CT and 20 MRI images from 40 patients. The CT and MRI images were obtained in different clinical centers. The cardiac structures of

the images were manually annotated by radiologists for both MRI and CT images. Our segmenter aimed to automatically segment four cardiac structures, including the ascending aorta (AA), the left atrium blood cavity (LA-blood), the left ventricle blood cavity (LV-blood), and the myocardium of the left ventricle (LV-myo). All the volumetric MRI and CT images and corresponding labels were preprocessed in [5], [6].

2) EVALUATION METRICS
We employed the Dice coefficient([%])to evaluate the agreement between the ground truth and predicted segmentation for cardiac structures. In addition, we calculated the average surface distance (ASD[voxel]) to measure the segmentation performance from the perspective of the boundary. A higher Dice and lower ASD reflect better segmentation performance. Both metrics are presented in the format of *mean±std*, which shows the average performance as well as the cross-subject variations of the results.

3) NETWORK ARCHITECTURES
In the segmenter, we use DRN101 [41] as the base semantic segmentation architecture, which is composed of stacked dilated residual blocks, to capture a large receptive field and the spatial context of feature maps. The DRN hyper-parameters used in our paper are the same as [6]. The multiple stages of layers [1, 3, 5, 7] are concatenated as image features **F**. The image feature dimension is $d = 512$. We modify the stride and dilation rate of the last layers followed by an upsampling layer to produce denser feature maps with a larger field of view for segmentation. The CDD consists of several stacked residual blocks followed by a sigmoid layer to predict the domain. The output dimension of the randomized multilinear map $d_o = 128$.

4) TRAINING DETAILS
The proposed model is implemented in PyTorch and trained on 4 NVIDIA RTX 2080 Ti GPUs with 44 GB memory. We train using the Adam optimizer [23] in three stages, and the batch size is 16, split equally for source and target samples. First, our model is trained from scratch on the source domain using a learning rate of $1e - 3$ to minimize the segmentation loss $\mathcal{L}_{seg}$. The trade-off parameter η in the $\mathcal{L}_{seg}$ (Eq. 2) item is 1. Second, the network is trained with only the source supervision and unsupervised domain classification losses $\mathcal{L}_{seg} + \lambda_1 \mathcal{L}_{cdd}$ at a learning rate of $3e - 4$. Then, the overall loss function (Eq. 13) is optimized, applying the conditional domain adversarial loss $\mathcal{L}_{cdd}$ and category-centric prototype alignment loss $\mathcal{L}_{ccpa}$ jointly, and the learning rate is set to $1e-4$. The trade-off ratio parameters of λ_1 and λ_2 are both set to the default standard setting value of 1 in the loss $\mathcal{L}_{ccpa}$. We also perform a series of experiments to analyze the impact of different hyper-parameters including λ_1, λ_2 and η, in the ablation study section. We use dropout with a ratio of 0.1 and batch normalization in all the convolutional layers. At the inference phase, the model directly

TABLE 1. Quantitative performance comparison between different methods on cardiac datasets (MRI → CT). (Note: **Bold** represent the best performance and the - denotes that the results were not reported by that method.)

Methods	AA		LA-blood		LV-blood		LV-myo	
	Dice	ASD	Dice	ASD	Dice	ASD	Dice	ASD
U-Net-MRI [52]	70.6±20.2	-	74.0±24.7	-	81.1±23.8	-	68.1±25.3	-
Cascaded-FCN-MRI [53]	76.6±13.8	-	81.1±13.8	-	87.7±7.7	-	75.2±12.1	-
PnP-AdaNet-v2-MRI [6]	80.3±1.8	-	78.1±8.0	-	88.3±2.1	-	70.8±2.1	-
U-Net-CT [52]	94.0±6.2	-	91.0±5.2	-	91.0±4.3	-	86.1±4.2	-
Cascaded-FCN-CT [53]	91.1±18.4	-	92.4±3.6	-	92.4±3.3	-	87.2±3.9	-
PnP-AdaNet-v2-CT [6]	77.6±29.2	-	89.4±2.2	-	91.3±2.5	-	85.8±1.5	-
Seg-MRI (ours)	78.3±5.5	15.3±10.6	78.9±7.1	15.7±6.4	88.5±5.2	2.9±0.5	75.3±3.1	2.3±2.1
Seg-CT (ours)	84.9±21.4	2.0±9.7	88.3±6.1	5.8±7.4	91.5±4.6	5.3±6.1	86.2±9.7	5.9±11.4
Seg-CT-noDA (ours)	32.3±8.4	14.4±19.3	28.9±14.3	8.9±3.8	2.9±1.1	19.2±9.4	15.7±13.1	28.4±33.9
DANN-CT-DA [54]	39.0±35.1	16.2±5.8	45.1±23.6	9.2±2.9	28.3±11.8	12.1±1.6	25.7±13.2	10.1±3.2
ADDA-CT-DA [28]	47.6±15.2	13.8±3.0	60.9±13.2	10.2±6.5	11.2±13.1	-	29.2±16.4	13.4±5.0
CycleGAN-CT-DA [29]	73.8±7.4	11.5±2.9	75.7±4.3	13.6±3.6	52.3±21.0	9.2±3.9	28.7±13.3	8.8±4.3
PnP-AdaNet-v2-CT-DA [6]	74.0±7.3	12.8±3.2	68.9±5.2	6.3±2.3	61.9±10.7	17.4±7.0	50.8±7.0	14.7±4.8
SIFA-v2-CT-DA [34]	81.3	7.9	**79.5**	6.2	73.8	**5.5**	61.6	8.5
3D based-CT-DA [55]	78.3±13.2	**2.7±3.2**	78.2±9.8	**3.8±1.2**	71.8±17.9	6.4±2.1	60.2±15.8	7.3±10.8
Seg-CT-DA (ours)	**81.8±18.7**	8.1±7.9	76.9±22.6	7.6±8.5	**75.9±15.8**	7.5±6.1	**62.3±16.1**	**7.1±6.9**

predicts every pixel that belongs to the most likely class via the segmenter on the source or target domain.

5) EXPERIMENTAL SETTINGS
To verify the performance of our proposed method, we conducted extensive experiments to demonstrate the performance of the domain adaptation method. In our experiments, biomedical cardiac MRI images were the source domain, and CT images were the target domain. In practice, we carry out the following experimental settings:

1) Training and testing the segmentation network only on the source domain (referred to Seg-MRI);

2) Training on the source domain and directly testing on target data, with no domain adaptation, as a lower bound (referred to Seg-CT-noDA);

3) Training and testing the segmentation network on annotated target domain images, as an upper bound (referred to Seg-CT);

4) Our method for unsupervised domain adaptation (referred to Seg-CT-DA).

B. EXPERIMENTAL RESULTS
We report segmentation quantitative results in this section, which demonstrate the superiority of the proposed method in the UDA scenario for cardiac structure segmentation, as illustrated in Table 1. In addition, Figure 4 presents the qualitative results of the segmentation for CT images.

First, we validated the performance of the segmenter for Seg-MRI, which serves as the basis for subsequent domain adaptation procedures. Our segmenter achieved competitive performance on most of the four cardiac structures compared with the standard U-Net [52] and cascaded-FCN [53]

methods. With the segmenter network architecture, we performed the following experiments to validate the effectiveness of our unsupervised domain adaptation framework.

Then we confirmed the upper-bounds performance of the segmenter on the target domain and report it as Seg-CT. Generally, these results are comparable to the standard U-Net [52] and cascaded-FCN [53] methods. Furthermore, comparing Seg-MRI with Seg-CT, we found a significant performance gap, which demonstrates the severe domain shift between the source and target domains. Furthermore, domain-shift problem inherent in cross-modality biomedical images is also illustrated by the degradation of Seg-CT-noDA's performance. This indicates that although the cardiac MRI and CT images share similar high-level representations and identical label space, the geometric pattern or boundary of each category or instance remains different, which makes domain adaptation extremely difficult.

What is striking about the figures in this table is that Seg-CT-DA outperforms Seg-CT-noDA in all metrics. Further analysis showed that the most striking aspect of the data was the largest increase in Dice of LV-blood. These results show the benefits of using a category-centric prototype to distinguish modal information in the cross-modality scenario, especially when neighboring structures contain unclear boundaries or relatively homogeneous tissues. Furthermore, as seen in the last part in Table 1, our method shows significant improvement over other domain adaptation methods in half of the metric. These results demonstrate the superiority of the proposed method in the UDA scenario for cardiac structure segmentation. In summary, these results show the importance of using CDD and CCPA across domains.

TABLE 2. Performance comparisons of the different components of our model on cardiac datasets (MRI → CT). The results indicate the importance of the proposed module for unsupervised domain adaptation. (Note: **Bold** represent the best performance.)

Methods	AA		LA-blood		LV-blood		LV-myo	
	Dice	ASD	Dice	ASD	Dice	ASD	Dice	ASD
Seg-CT-DA (Standard Setting: $\lambda_1 = 1, \lambda_2 = 1, \eta = 1$)	**81.8±18.7**	**8.1±7.9**	**76.9±22.6**	**7.6±8.5**	**75.9±15.8**	**7.5±6.1**	**62.3±16.1**	**7.1±6.9**
w/o CDD ($\lambda_1 = 0$)	73.6±16.1	11.7±14.4	68.2±9.2	8.1±6.1	60.1±8.6	9.8 ±7.5	49.6±7.4	15.4±15.3
w/o CCPA ($\lambda_2 = 0$)	74.9±13.7	12.5±14.3	66.2±18.6	7.9±4.3	57.8±3.5	9.5 ±9.6	48.2±14.7	14.3±24.7
w/o CDD & CCPA	32.3±8.4	14.4±19.3	28.9±14.3	8.9±3.8	2.9±1.1	19.2 ±9.4	15.7±13.1	28.4±33.9
$\lambda_1 = 0.3$	75.4±31.5	10.4±8.9	71.9±12.8	8.0±5.6	64.8±12.5	8.6±5.3	53.8±30.2	13.1±3.7
$\lambda_1 = 0.6$	78.5±28.2	9.7±8.3	75.3±3.9	7.8±4.3	68.3±32.0	8.0±9.3	56.9±21.5	10.4±3.3
$\lambda_1 = 0.9$	80.3±31.8	8.8±5.3	76.1±10.2	7.7±5.2	73.2±21.7	7.8±7.1	61.4±22.7	8.2±7.2
$\lambda_2 = 0.3$	77.3±12.5	10.2±4.2	70.5±5.3	7.9±2.6	65.3±22.5	8.9±6.2	52.5±31.7	12.5±4.2
$\lambda_2 = 0.6$	79.6±22.7	9.4±5.4	72.3±2.5	7.8±3.5	70.8±11.8	8.4±2.5	56.2±22.8	9.3±2.4
$\lambda_2 = 0.9$	81.1±6.4	8.6±7.1	75.1±4.9	7.7±4.8	74.3±26.4	7.6±2.8	61.9±18.1	7.8±3.7
$\eta = 0$	76.0±18.2	9.6±5.7	70.1±1.9	8.5±3.2	67.2±12.4	8.7±4.9	54.8±12.7	11.7±8.2
$\eta = 0.2$	78.3±24.6	9.2±3.3	71.5±3.5	8.2±4.5	69.6±23.5	8.4±2.4	56.7±13.6	10.1±3.0
$\eta = 0.4$	79.6±14.9	8.8±6.7	73.4±8.4	8.0±5.4	71.5±31.6	8.0±4.1	58.1±17.6	9.4±3.6
$\eta = 0.6$	80.6±6.8	8.6±2.4	74.2±4.6	7.9±4.1	72.8±21.8	7.8±2.9	60.8±11.9	8.6±4.1
$\eta = 0.8$	81.4±24.7	8.3±2.1	75.6±8.9	7.7±2.9	74.0±14.3	7.7±6.4	61.3±17.5	7.4±5.3

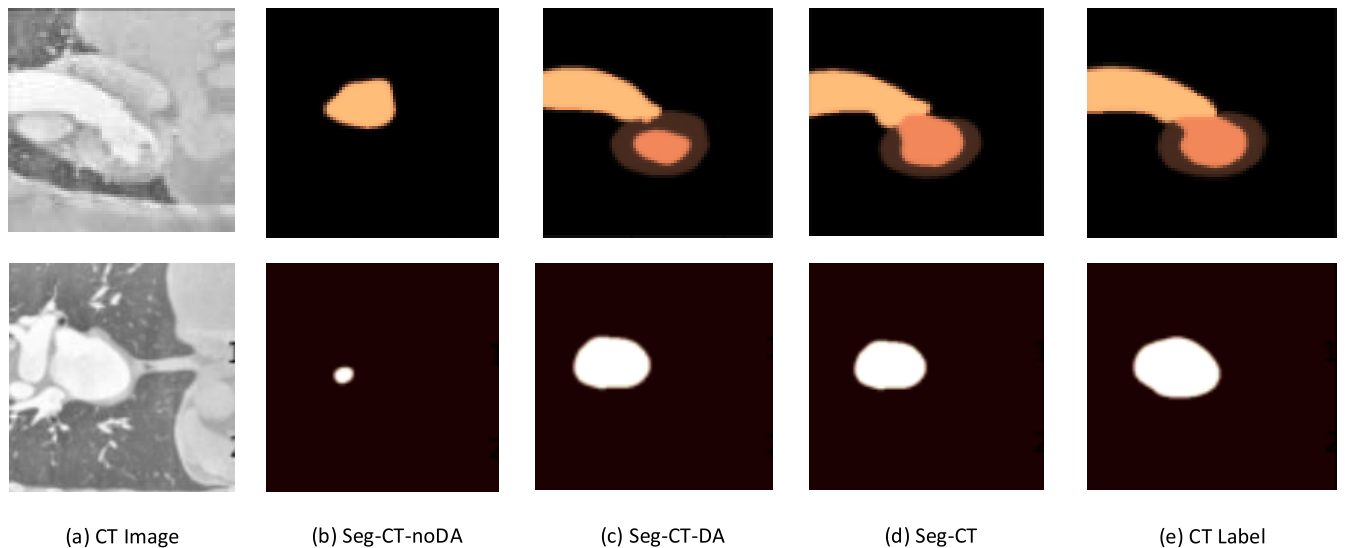| (a) CT Image | (b) Seg-CT-noDA | (c) Seg-CT-DA | (d) Seg-CT | (e) CT Label |

FIGURE 4. Results of different methods for CT image segmentations. Each row presents one typical example, from left to right: (a) raw CT image, (b) directly applying the MRI segmenter on CT data, (c) our unsupervised cross-modality domain adaptation result, (d) the segmenter trained from scratch with CT labels, and (e) ground truth labels. The structures of AA, LA-blood, LV-blood and LV-myo are indicated by brown, white, red and light red colors, respectively.

C. ABLATION STUDIES

1) INFLUENCE OF KEY COMPONENT

To evaluate the contributions of each component of our model, we perform ablation studies in this section. As described in Table 2, when we remove either the CDD module ($\lambda_1 = 0$) or CCPA module ($\lambda_2 = 0$) from our method, the most striking observation to emerge from the data comparison is the decrease in performance on all metrics. This indicates that both two modules play an important role in addressing the issue of domain shift in UDA. Furthermore,

when we remove these two modules ($\lambda_1 = 0, \lambda_2 = 0$), significant performance degradation occurs. This shows that enabling domain adaptation helps to improve the model generalization capability on cross-modality segmentations.

2) INFLUENCE OF KEY HYPER-PARAMETERS

We also perform a series of ablation studies to analyze the impact of different hyper-parameters on the domain adaptation performance. All models are trained from scratch with the default standard settings. Here, we study three key

TABLE 3. Quantitative performance of reverting the domain adaptation direction comparison between different methods on cardiac datasets (CT → MRI). (Note: **Bold** represent the best performance.)

Methods	AA		LA-blood		LV-blood		LV-myo	
	Dice	ASD	Dice	ASD	Dice	ASD	Dice	ASD
Seg-MRI (ours)	78.3±5.5	15.3±10.6	78.9±7.1	15.7±6.4	88.5±5.2	2.9±0.5	75.3±3.1	2.3±2.1
Seg-MRI-noDA (ours)	34.8±13.8	20.7±13.6	10.6±13.8	19.0±8.1	5.8±4.6	11.2 ±3.6	16.4±11.7	6.4±3.9
AdaOutput-MRI-DA [42]	60.8	**5.7**	39.8	8.0	71.5	4.6	35.5	4.6
PnP-AdaNet-v2-MRI-DA [6]	43.7±10.8	11.4±3.2	47.0±7.3	14.5±4.1	77.7±10.4	4.5±1.4	**48.6±2.9**	5.3±1.8
SIFA-v1-MRI-DA [33]	**67.0**	6.2	60.7	9.8	75.1	4.4	45.8	4.4
SIFA-v2-MRI-DA [34]	65.3	7.3	62.3	**7.4**	78.9	3.8	47.3	4.4
Seg-MRI-DA (ours)	66.4±28.3	16.9±10.8	**63.0±34.2**	17.1±12.9	**79.2±12.6**	**3.5±4.4**	47.8±8.4	**4.1±7.3**

hyper-parameters, the trade-off ratio λ_1 and λ_2 of $\mathcal{L}_{cdd}$ and $\mathcal{L}_{ccpa}$ in Eq. 13, respectively, and the trade-off ratio η of the Dice loss in Eq. 2. The results are shown in Table 2, consisting of three groups of experimental comparisons at the bottom of Table 2.

Fixing $\lambda_2 = 1, \eta = 1$, we vary λ_1 ranging in [0.3, 0.6, 0.9, 1]. We first observe that as the value of λ_1 increases, the domain adaptation performance increases gradually. Compared to $\lambda_1 = 0.3$, $\lambda_1 = 1$ obtains performance improvement on all metrics, especially on the Dice of LV-blood, and the ASD of LV-myo. These tissues are difficult and contain unclear boundaries. We believe this phenomenon is because the introduced CDD module can deal well with cross-modality domain adaptation. In practice, we should set a task-specific value of λ_1 ensuring optimal results.

Similarly, we evaluate different settings $\lambda_2 = [0.3, 0.6, 0.9, 1]$ of the trade-off ratio of $\mathcal{L}_{ccpa}$ in the Eq. 13. As an interesting observation, the change in the metrics is monotonic as we adjust the trade-off ratio. Additionally, $\lambda_2 = 1$ likewise achieves the best performance on all metrics, in particular, a significant improvement in the Dice of LV-blood and the ASD of LV-myo. This indicates the importance of the CCPA module in cross-modality domain adaptation segmentation.

Finally, we perform a group of experiments on the hyper-parameters η to investigate how it affects the model performance. We report the results at the bottom of Table 2 by adjusting η ranging from [0, 0.2, 0.4, 0.6, 0.8, 1]. We observe in Table 2 that our method achieves the best performance under the setting $\eta = 1$ compared with other settings. Note that the cardiac structure datasets exhibit class imbalance. In this regard, it is beneficial for challenging heart segmentation tasks to combine Dice loss with cross-entropy loss. In practice, we are supposed to adjust the optimal ratio η according to the degree of sample class imbalance in the dataset.

3) REVERTING DOMAIN ADAPTATION DIRECTION
To investigate whether the reverse adaptation direction from CT to MRI can also be achieved and what the impact of

$\mathcal{D}_s$ in domain adaptation is, we applied the same model setup, replacing the source domain as CT and the target domain as MRI for the experiments. The quantitative performance of reverting the domain adaptation direction is shown in Table 3. Unsurprisingly, using the CT segmenter directly on MRI data also fails. Our proposed method can recover the average segmentation performances via unsupervised domain adaptation, indicating that our model has good robustness and generalization over different types of datasets and that cross-modality domain adaptation can be achieved in both directions. Interestingly, the best recovered structure in this reverse setup is the LV-blood, which is the same structure as the MRI-to-CT direction. It is also necessary to mention that the Dice of LV-blood increases from a complete failure (5.8) to a considerably higher value (79.2), and the ASD of LV-blood also shows the same trend. Compared to the adaptation from MRI to CT, the reverse direction generally yields lower performance generally, show that the difficulty of $\mathcal{D}_s$ can affect the performance of domain adaptation, and the difficulty is not symmetric. Segmentation of cardiac MRI itself is more difficult than segmentation of cardiac CT. This is also evident from Table 1, where CT segmentation Dice is higher than the result of MRI in all four structures. In these respects, transferring CT segmenter to MRI seems to be more challenging.

V. CONCLUSION
In this paper, we studied the problem of unsupervised domain adaptation on cross-modality biomedical images. We introduced the CDD and CCPA modules into the model to align the latent feature space and category-centric prototype of the target domain to that of the source domain. Extensive experiments show the superior performance of our approach.

REFERENCES
[1] S. A. Taghanaki, K. Abhishek, J. P. Cohen, J. Cohen-Adad, and G. Hamarneh, "Deep semantic segmentation of natural and medical images: A review," *Artif. Intell. Rev.*, vol. 54, no. 1, pp. 137–178, 2020.

[2] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[3] O. Çiçek, A. Abdulkadir, S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2016, pp. 424–432

[4] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Mach. Learn.*, vol. 79, nos. 1–2, pp. 151–175, May 2010.

[5] 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.

[6] Q. Dou, C. Ouyang, C. Chen, H. Chen, B. Glocker, X. Zhuang, and P.-A. Heng, "PnP-AdaNet: Plug-and-play adversarial domain adaptation network at unpaired cross-modality cardiac segmentation," *IEEE Access*, vol. 7, pp. 99065–99076, 2019.

[7] H. Yan, Y. Ding, P. Li, Q. Wang, Y. Xu, and W. Zuo, "Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 945–954.

[8] M. Long, Y. Cao, J. Wang, and M. I. Jordan, "Learning transferable features with deep adaptation networks," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 97–105.

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

[10] Z. Pei, Z. Cao, M. Long, and J. Wang, "Multi-adversarial domain adaptation," in *Proc. AAAI Conf. Artif. Intell.* Menlo Park, CA, USA: Association for the Advancement of Artificial Intelligence, 2018, pp. 3934–3941.

[11] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *J. Mach. Learn. Res.*, vol. 13, pp. 723–773, Mar. 2012.

[12] S. Sankaranarayanan, Y. Balaji, C. D. Castillo, and R. Chellappa, "Generate to adapt: Aligning domains using generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8503–8512.

[13] M. Xu, H. Wang, B. Ni, Q. Tian, and W. Zhang, "Cross-domain detection via graph-induced prototype alignment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12352–12361.

[14] Y. Pan, T. Yao, Y. Li, Y. Wang, C.-W. Ngo, and T. Mei, "Transferrable prototypical networks for unsupervised domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2234–2242.

[15] K. Saito, D. Kim, S. Sclaroff, T. Darrell, and K. Saenko, "Semi-supervised domain adaptation via minimax entropy," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8049–8057.

[16] S. Xie, Z. Zheng, L. Chen, and C. Chen, "Learning semantic representations for unsupervised domain adaptation," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 5423–5432.

[17] M. Long, Z. Cao, J. Wang, and M. I. Jordan, "Conditional adversarial domain adaptation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 1640–1650.

[18] M. Toldo, A. Maracani, U. Michieli, and P. Zanuttigh, "Unsupervised domain adaptation in semantic segmentation: A review," 2020, *arXiv:2005.10876*. [Online]. Available: http://arxiv.org/abs/2005.10876

[19] S. Sun, H. Shi, and Y. Wu, "A survey of multi-source domain adaptation," *Inf. Fusion*, vol. 24, pp. 84–92, Jul. 2015.

[20] M. Wang and W. Deng, "Deep visual domain adaptation: A survey," *Neurocomputing*, vol. 312, pp. 135–153, Oct. 2018.

[21] V. M. Patel, R. Gopalan, R. Li, and R. Chellappa, "Visual domain adaptation: A survey of recent advances," *IEEE Signal Process. Mag.*, vol. 32, no. 3, pp. 53–69, May 2015.

[22] C. Chen, C. Qin, H. Qiu, G. Tarroni, J. Duan, W. Bai, and D. Rueckert, "Deep learning for cardiac image segmentation: A review," *Frontiers Cardiovascular Med.*, vol. 7, p. 25, Mar. 2020.

[23] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," 2014, *arXiv:1412.3474*. [Online]. Available: http://arxiv.org/abs/1412.3474

[24] J. Shen, Y. Qu, W. Zhang, and Y. Yu, "Wasserstein distance guided representation learning for domain adaptation," in *Proc. AAAI Conf. Artif. Intell.* Menlo Park, CA, USA: Association for the Advancement of Artificial Intelligence, 2017, pp. 1–8.

[25] Y. Ding, H. Fan, M. Xu, and Y. Yang, "Adaptive exploration for unsupervised person re-identification," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 16, no. 1, pp. 1–19, Apr. 2020.

[26] H. Fan, L. Zheng, C. Yan, and Y. Yang, "Unsupervised person re-identification: Clustering and fine-tuning," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 14, no. 4, pp. 1–18, Nov. 2018.

[27] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 2, pp. 2672–2680, 2014.

[28] 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. 2962–2971.

[29] 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. (ICCV)*, Oct. 2017, pp. 2223–2232.

[30] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. A. Efros, and T. Darrell, "CyCADA: Cycle-consistent adversarial domain adaptation," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1989–1998.

[31] P. Russo, F. M. Carlucci, T. Tommasi, and B. Caputo, "From source to target and back: Symmetric bi-directional adaptive GAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8099–8108.

[32] Y. Zhang, S. Miao, T. Mansi, and R. Liao, "Unsupervised X-ray image segmentation with task driven generative adversarial networks," *Med. Image Anal.*, vol. 62, May 2020, Art. no. 101664.

[33] C. Chen, Q. Dou, H. Chen, J. Qin, and P. Heng, "Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33. Menlo Park, CA, USA: Association for the Advancement of Artificial Intelligence, 2019, pp. 865–872.

[34] C. Chen, Q. Dou, H. Chen, J. Qin, and P. A. Heng, "Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 7, pp. 2494–2505, Jul. 2020.

[35] M. Ghafoorian, A. Mehrtash, T. Kapur, N. Karssemeijer, E. Marchiori, M. Pesteie, C. Guttmann, F. Leeuw, C. Tempany, B. Ginneken, A. Fedorov, P. Abolmaesumi, B. Platel, and W. Wells, "Transfer learning for domain adaptation in MRI: Application in brain lesion segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2017, pp. 516–524.

[36] A. van Opbroek, M. A. Ikram, M. W. Vernooij, and M. de Bruijne, "Transfer learning improves supervised image segmentation across imaging protocols," *IEEE Trans. Med. Imag.*, vol. 34, no. 5, pp. 1018–1030, May 2015.

[37] 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. Int. Conf. Inf. Process. Med. Imag.* Cham, Switzerland: Springer, 2017, pp. 597–609.

[38] M. A. Degel, N. Navab, and S. Albarqouni, "Domain and geometry agnostic CNNs for left atrium segmentation in 3D ultrasound," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2018, pp. 630–637.

[39] L. Zhang, M. Pereañez, S. Piechnik, S. Neubauer, S. Petersen, and A. Frangi, "Multi-input and dataset-invariant adversarial learning (MDAL) for left and right-ventricular coverage estimation in cardiac MRI," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2018, pp. 481–489.

[40] S. Wang, L. Yu, X. Yang, C.-W. Fu, and P.-A. Heng, "Patch-based output space adversarial learning for joint optic disc and cup segmentation," *IEEE Trans. Med. Imag.*, vol. 38, no. 11, pp. 2485–2495, Nov. 2019.

[41] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 636–644.

[42] Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker, "Learning to adapt structured output space for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7472–7481.

[43] Y.-H. Chen, W.-Y. Chen, Y.-T. Chen, B.-C. Tsai, Y.-C.-F. Wang, and M. Sun, "No more discrimination: Cross city adaptation of road scene segmenters," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2011–2020.

[44] P. Kar and H. Karnick, "Random feature maps for dot product kernels," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2012, pp. 583–591.

[45] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 1177–1184.

[46] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 214–223.

[47] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.

[48] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–12.

[49] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[50] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Perez, "ADVENT: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2512–2521.

[51] 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.

[52] 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.*, 2015, pp. 234–241.

[53] C. Payer, D. Stern, H. Bischof, and M. Urschler, "Multi-label whole heart segmentation using CNNS and anatomical label configurations," in *Proc. Int. Workshop Stat. Atlases Comput. Models Heart, (MICCAI)*, 2017, pp. 190–198.

[54] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2030–2096, 2016.

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

WENWEN YU received the B.S. degree from the School of Medical Information and Engineering, Xuzhou Medical University (XZHMU), Xuzhou, China, in 2018. He is currently pursuing the master's degree with the School of Medical Imaging, XZHMU. His research interests include computer vision, machine learning, in particular, the area of medical image analysis and document understanding.

QIUWEN SUN received the B.S. degree from the School of Medical Information and Engineering, Xuzhou Medical University (XZHMU), Xuzhou, China, in 2019. She is currently pursuing the master's degree with the School of Medical Imaging, XZHMU. Her research interest includes bioinformatics.

RUOHAN ZHAO received the B.S. degree from the Beijing University of Chemical Technology, Beijing, China, in 2014, and the M.S. degree from The Hong Kong Polytechnic University (PolyU), Hong Kong, in 2016, where he is currently pursuing the Ph.D. degree. His main research interest includes medical image processing.

PING GONG received the B.S., M.S., and Ph.D. degrees from the China University of Mining and Technology (CUMT), Xuzhou, China, in 2003, 2006, and 2019, respectively. She is currently an Associate Professor with the School of Medical Imaging, Xuzhou Medical University (XZHMU), Xuzhou. Her main research interests include medical image processing and bioinformatics.

JUNFENG HU received the Ph.D. degree from the China University of Mining and Technology (CUMT), Xuzhou, China, in 2010. He is currently a Professor with the School of Medical Imaging, Xuzhou Medical University (XZHMU), Xuzhou. His main research interest includes medical image processing.

● ● ●