Article:
Vesal, S, Gu, M, Kosti, R et al. (2 more authors) (2021) Adapt Everywhere: Unsupervised Adaptation of Point-Clouds and Entropy Minimisation for Multi-modal Cardiac Image Segmentation. IEEE Transactions on Medical Imaging. ISSN 0278-0062

https://doi.org/10.1109/tmi.2021.3066683

Adapt Everywhere: Unsupervised Adaptation of Point-Clouds and Entropy Minimisation for Multi-modal Cardiac Image Segmentation

Sulaiman Vesal, Mingxuan Gu, Ronak Kosti, Andreas Maier *Member, IEEE*, and Nishant Ravikumar

***Abstract*— Deep learning models are sensitive to domain shift phenomena. A model trained on images from one domain cannot generalise well when tested on images from a different domain, despite capturing similar anatomical structures. It is mainly because the data distribution between the two domains is different. Moreover, creating annotation for every new modality is a tedious and time-consuming task, which also suffers from high inter- and intra- observer variability. Unsupervised domain adaptation (UDA) methods intend to reduce the gap between source and target domains by leveraging source domain labelled data to generate labels for the target domain. However, current state-of-the-art (SOTA) UDA methods demonstrate degraded performance when there is insufficient data in source and target domains. In this paper, we present a novel UDA method for multi-modal cardiac image segmentation. The proposed method is based on adversarial learning and adapts network features between source and target domain in different spaces. The paper introduces an end-to-end framework that integrates: a) entropy minimisation, b) output feature space alignment and c) a novel point-cloud shape adaptation based on the latent features learned by the segmentation model. We validated our method on two cardiac datasets by adapting from the annotated source domain, bSSFP-MRI (balanced Steady-State Free Procession-MRI), to the unannotated target domain, LGE-MRI (Late-gadolinium enhance-MRI), for the multi-sequence dataset; and from MRI (source) to CT (target) for the cross-modality dataset. The results highlighted that by enforcing adversarial learning in different parts of the network, the proposed method delivered promising performance, compared to other SOTA methods.**

***Index Terms*— Unsupervised Domain Adaptation, Cardiac Segmentation, multi-modal Segmentation, Adversarial Learning, Point-Clouds, Entropy Minimisation**

I. Introduction

MYOCARDIAL infarction (MI) is a cardiovascular disease with a high percentage of mortality and morbidity rate worldwide [1,2]. For the diagnosis and treatment of patients with MI, there is a need for modelling of the ventricle blood pools and accurate analysis of myocardium using different imaging modalities [3]. Cardiac magnetic resonance (CMR) imaging modalities are regularly used in the clinical diagnosis to provide anatomical morphology and operational information of the heart. A comprehensive diagnosis involves different types of CMR sequences which provide complementary information to each other. Among them LGE images are desirable to determine the presence, location, and extent of MI [4]. However, manual contouring is usually time-consuming, tiresome, and subjected to inter- and intra-observer variations [5]. Due to the generation of multiple imaging modalities, there is a substantial clinical need to develop a multi-modal CMR segmentation system that can generalise well across different modalities [6].

Recently, Convolutional Neural Network (CNN) based methods have been widely used for medical image analysis in detection, segmentation [7,8], and tracking of anatomical structures. Such methods are often generic and can be extended from one imaging modality to another by fine-tuning or re-training on the target imaging modality. However, to achieve satisfactory performance, a sufficient number of annotated target training images are required. In practice, it is often difficult to accumulate enough training images for a new imaging modality not well established in clinical practice yet. Synthesising or data augmentation is often used to support training data in the hope that they can boost the generalisation capability of a trained deep learning model. However, the distribution gap between synthesised data and real data often determines the success of such an approach.

Taking advantage of unlabelled data from other modalities is quite challenging due to the high data distribution discrepancy. Recent advances have used generative adversarial networks (GAN) [9] to formulate it as an image-to-image translation task. These methods require pixel-to-pixel correspondence between two domains to build a direct cross-modality segmentation model. However, multi-modal medical images are mostly in 3D and do not have cross-modal paired data. A method to learn from unpaired data is clinically more desirable. Anatomical shape in medical images and volumes contain diagnostic information, so it is important to maintain translation invariance. However, GAN models that are trained without paired data do not guarantee this requirement

S. Vesal, M. Gu, R.Kosti, and A. Maier are with the Pattern Recognition Lab, Friedrich-Alexander-University Erlangen-Nuremberg, Germany. (E-mail: sulaiman.vesal@fau.de)

N. Ravikumar is with CISTIB, Centre for Computational Imaging and Simulation Technologies in Biomedicine, School of Computing, LICAMM Leeds Institute of Cardiovascular and Metabolic Medicine, School of Medicine, University of Leeds, United Kingdom.

The work described in this paper was partially supported by the project EFI-BIG-THERA: Integrative 'BigData Modeling' for the development of novel therapeutic approaches for breast cancer. The authors would also like to thank NVIDIA for donating a Titan X-Pascal GPU.

due to the lack of direct reconstruction and the reliance on discriminators. Therefore, adapting features between different domains can avoid any complex mapping process between paired images [10,11].

In this work, we propose a new scheme to leverage entropy and shape information available in the source and target domain for UDA. We hypothesise that introducing additional shape information using point-clouds along with entropy adaptation brings complementary effects to further bridge the performance gap between source and target at test time. To this end, we transform the segmentation network such that the shape information in the form of point-clouds is embedded into a dedicated deep architecture using an auxiliary point-cloud regression task. Point-clouds for cardiac shape estimation, operating as an additional source domain supervision in our framework (only available while training), will be considered as furnished information. Another challenge is to incorporate 2D point-cloud regression into UDA learning efficiently. We achieve this by introducing a new point-cloud adversarial training protocol based on PointNet discriminator. The proposed approach permits accurate segmentation of cardiac images without any annotation.

In summary, our main contributions are threefold:

- *First*, we propose a novel UDA method based on adversarial learning that adapts the features between source and target domains in different spaces. Our network incorporates entropy minimisation, output space alignment, and point-cloud adaptation to tackle drastic domain shift. To the best of our knowledge, it is the first study that employed shape-prior information using point-clouds for UDA.
- *Second*, we present a novel multi-task model for point-set generation from the latent representation of the segmentation network. It allows multi-task learning with point-set network acting as a surface shape learning mechanism and consequently improves the segmentation network for the new domain.
- *Third*, we validate the proposed point-cloud UDA on two cardiac image segmentation tasks, including multi-sequences CMR (bSSFP, T2-weighted, and LGE) and cross-modal cardiac CT and MRI. The proposed method achieved promising results. Code and models available at: `https://github.com/sulaimanvesal/PointCloudUDA`

While working with multi-sequence data, MS-CMRSeg [12], the source and target domains are bSSFP-MRI (we combine bSSFP and T2 sequences into source domain) and LGE-MRI; whereas for cross-modal data, MM-WHS [13], source and target domains are MRI and CT, respectively.

The rest of the paper is organised as follows: Section II reviews related work; Section III describes our proposed entropy and shape-aware UDA method; Section IV deals with comprehensive evaluation and analysis of our experiments; Section V discusses the statistical analysis and limitation of the proposed method, and Section VI concludes our work.

II. Related Work

Recently, UDA methods have been explored in the field of medical imaging to deal with the performance degradation caused by the domain shift. Existing methods on UDA generally suggest aligning the source and target domain distributions from three perspectives. First category is the image-level alignment, which transforms the image appearance between domains with an image-to-image transformation model [10,14]–[18]. The second category focuses on feature alignment, aiming to extract domain-invariant features usually by minimising feature distance between domains via adversarial learning [19]–[21]. And, a recent third category focuses on alignment of feature-level and image-level information [7].

Image-level domain adaptation: Image-based UDA methods are developed based on unpaired image-to-image translation algorithms, which mainly use CycleGAN [14] and MUNIT [22]. Bousmalis *et al.* [23] aligned the image appearance between two modalities using the pixel-to-pixel transformation, employing domain adaptation at input space. In such models, domain adaptation relies highly on the quality of stylised images, which are often not perfect. Zhang *et al.* [24] used cycle and shape-consistency adversarial networks for multi-modal brain MRI segmentation. In [8], the authors again used image translation to synthesise the data and incorporate it with an attention-based neural network. Delisle *et al.* [25] developed an adversarial method to tackle UDA segmentation from a normalisation perspective.

Feature-level domain adaptation: In the feature-level, Kamnitsas *et al.* [26] proposed a UDA for brain lesion segmentation, which learned domain-invariant features with a discriminator that predicts the input image domain. In cross-modal segmentation with drastic differences between the source and target domain, Q. Dou *et al.* [27] fine-tuned specific feature layers, and employed adversarial loss for supervised feature learning. Recently, output space alignment is also exploited to incorporate the spatial and structural geometry information of predictions [19]. Wang *et al.* [28] proposed a method to adversarially adapt entropy and boundary of Fundus images of different vendors. The new depth-aware adaptation scheme, DADA learning [29], simultaneously aligns segmentation and depth-based information of source and target while being aware of scene geometry.

Image and Feature domain adaptation: CyCADA [11] poses UDA as style transfer with adversarial learning to bridge the gap in appearance between the source and target domains, while simultaneously aligning the image and latent feature spaces independently. To avoid divergence of semantics, they enforce cycle consistency during adaptation of the corresponding domains. Chen *et al.* [7] proposed a domain adaptation framework, SIFA, which considers both feature and image-level adaptation, concurrently for MRI to CT segmentation. Their network is built upon a CycleGAN, with additional discriminators to emphasise on feature-level adaptation, and image domain separation. They extended SIFA to Bidirectional-SIFA [30] by adding deeply supervised feature alignment and exploring adaptation in a bi-directional fashion ($MRI \Leftrightarrow CT$) achieving SOTA on multi-modal medical

image segmentation.

Multi-sequence domain adaptation: For multi-sequence cardiac MRI segmentation, Chen *et al.* [10] proposed a network (bSSFP $\rightarrow$ LGE) based on multi-modal unsupervised image-to-image translation (MUNIT) network [22]. The method transferred style, shape, and appearance from bSSFP images to LGE to generate synthetic samples, to train the segmentation network. Wang *et al.* [31] proposed a fully end-to-end unsupervised method based on adversarial training to minimise discrepancies in both the feature and output space. Roth *et al.* [32] couples classical methods of multi-atlas label fusion with deep learning by formulating noisy labels for unlabelled LGE images using the registration technique.

Most of the previous methods fail to produce reliable predictions when the target images are noisy, visually different, or without clear boundaries. Cai *et al.* [33] proposed using point-clouds to incorporate shape information during training, where their segmentation model is aware of the shape and topology of organs. Their shape learning multi-task network uses multi-scale features from the segmentation model to generate organ surface point-clouds with more details. They also incorporated a discriminator to improve the generation of point-clouds with fewer outliers in an adversarial fashion. As a result, the shape learning step acts as an auxiliary task to provide complementary information for the segmentation model and improves organ segmentation. Clearly, the shape information for cardiac ventricle segmentation is also an essential factor. Therefore, developing an effective UDA method to improve the prediction performance on the shape of the overall ventricles of the target domain images remains a challenge.

III. Entropy and Shape-Aware Domain Adaptation

Problem Definition: In UDA for semantic segmentation, we are given a set of source images, bSSFP-MRI (multi-sequence data) or MRI (multi-modal data), and their corresponding mask labels in the source domain $\mathbb{D}_s = (\mathbf{I}_i^s, \mathbf{Y}_i^s)_{i=1}^{m_s}$, where $\mathbf{I}_i^s \in \mathcal{R}^{w \times h \times 3}$, $\mathbf{Y}_i^s \in \mathcal{R}^{w \times h \times c}$ and m_s is the number of source images. In the target domain, LGE-MRI (multi-sequence data) or CT (multi-modal data), we are given unlabelled target images $\mathbb{D}_t = (\mathbf{I}_i^t)_{i=1}^{m_t}$, where m_t is the number of target images. Our goal is to train a supervised model on $\mathbb{D}_s$ and incorporate information from $\mathbb{D}_t$ to reduce the gap between two domains, and improve segmentation accuracy on $\mathbb{D}_t$ in an unsupervised manner. Due to domain shift, images across domains usually present different data distribution; the goal is to bring distributions of both the domains closer.

A. Overview of the Proposed Method

We hypothesise that offering additional shape information using point-clouds along with feature adaptation brings complementary effects to bridge the performance gap between source and target at test time. Starting from an existing segmentation network, we insert additional modules: *(1)* to predict object point-clouds as supplementary output, and *(2)* to feed the information exploited by this auxiliary task back to the main network. To align the features for $\mathbb{D}_s$ and $\mathbb{D}_t$, the most common way is applying adversarial learning directly in feature space, such that a discriminator learns to differentiate the domain space of the features. However, due to the high dimensionality of the feature space, it is difficult to align the features directly. Instead, we enhance the domain-invariance of feature distributions by: *(1)* using adversarial learning via entropy minimisation, *(2)* aligning objects' shape information with point-cloud data, and *(3)* discriminating the structured output space.

Our proposed UDA framework consists of a multi-task segmentation and point-cloud regression network $\mathbf{G}$, which has shared weights between $\mathbb{D}_s$ and $\mathbb{D}_t$. We modified DR-UNet [34] segmentation network by adding another head to the encoder part. This head estimates the latent representation features of the encoder to a vector of size 300×3, which is the shape information of the input image in the form of point-cloud. We constructed three discriminators *viz.* $\mathbf{D_3}$ for point-cloud surface shape alignment, $\mathbf{D_1}$ for segmentation output adaptation and $\mathbf{D_2}$, which receives self-weighted information map for entropy minimisation.

During training, we first provide the images $\mathbf{I}_s \in \mathcal{R}^{h \times w \times 3}$ (with annotations) from source domain to the segmentation network in a supervised manner to optimise $\mathbf{G}$. We then take the images from the target domain and predict the segmentation output $\mathbf{S}_t \in \mathcal{R}^{h \times w \times 4}$. Once we have the output probability maps for both domains, we convert them to self-weighted information to compute entropy. Then, the network back-propagates gradients from $\mathbf{D}_1$, $\mathbf{D}_2$, and $\mathbf{D}_3$ to $\mathbf{G}$, which encourages $\mathbf{G}$ to optimise its weights w.r.t the segmentation labels from source and target domains. Fig. 1 shows an overview of the proposed network. Our pipeline takes advantage of entropy minimisation and output space alignment, so we briefly review these methods before introducing our shape-aware point-cloud based UDA method.

Multi-task Segmentation and Point-Cloud Regression Network: Our segmentation network (Fig. 2), $\mathbf{G}$, has an encoder that extracts high-level features from input images. The encoder has four levels and a bottleneck similar to DR-UNet [34], which also includes residual connections and dilated convolutions in the last level. We constructed two independent decoder heads for output segmentation and point-cloud regression. The decoder for segmentation is also similar to the DR-UNet network. It consists of four levels of 2D convolution, nearest neighbour up-sampling, and a 1×1 convolution. The point-cloud regression head has one convolution with a kernel size of 6×6 and a fully connected layer, which takes the encoder latent features and generates a shape vector (300×3) of the point-cloud.

We use PatchGAN [9,35] for the discriminators $\mathbf{D}_1$ and $\mathbf{D}_2$, which takes two inputs (e.g. source and target probability maps or entropy maps) and distinguishes patches of size 8×8. PatchGAN discriminator mainly penalises structure at the scale of local image patches and is run convolutionally across the input images. There are five convolutional layers in this architecture, with a kernel size of 4×4 and a stride shape of 2, and the number of feature maps for each layer is: $[64, 128, 256, 512, 1]$, respectively. Each convolution layer is followed by a leaky ReLU activation map, parameterised with 0.2.

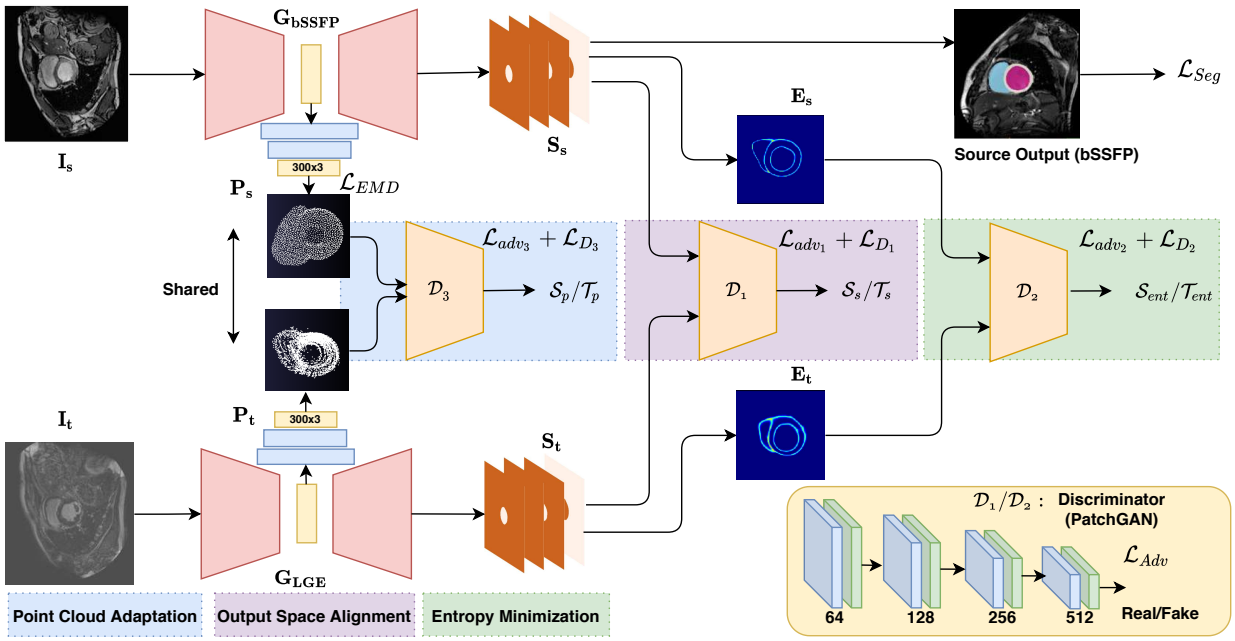


Fig. 1. Overview of the proposed UDA segmentation framework. A single image from the source domain or the target domain fed into its corresponding domain-specific DR-UNet segmentation network (depicted in Fig. 2), which has shared weights. The DR-UNet encoder extracts high-level features for both the source and target domains. Then the features are sent to the decoder for segmentation and point-net to generate point-cloud. The point-cloud is fed to $\mathbf{D_3}$ for shape alignment (depicted in Fig. 3). The output probability of DR-UNet for the source domain is trained in a supervised manner. Then, we send the *softmax* output simultaneously to output-space alignment and entropy minimisation discriminators. The domain classifier networks, $\mathbf{D_1}$ and $\mathbf{D_2}$, then differentiates whether its input is from source domain or target domain.

B. Output Feature Space Alignment

The first component of our network mainly focuses on adapting the distribution of output probability segmentation masks based on the adversarial learning approach similar to [19]. The segmentation network $\mathbf{G}$ predicts output segmentation masks, and for adversarial learning, a discriminator is included to recognise whether the input image is from the source or target domains. Here, we denote images from source and target domains as $\mathbf{I_s}$, $\mathbf{I_t} \in \mathcal{R}^{h \times w \times 3}$. The segmentation network $\mathbf{G}$ first receives the source images $\mathbf{I_s}$ for supervised learning and predicts the segmentation output $\mathbf{S}_s$. Then we predict the segmentation output $\mathbf{S}_t$ for the target image $\mathbf{I_t}$ (without annotations). With an aim to align $\mathbf{S}_s$ and $\mathbf{S}_t$ closer to each other, $\mathbf{D_1}$ takes these predictions as its input to differentiate whether it is from $\mathbb{D}_s$ or $\mathbb{D}_t$. The network propagates gradients with an adversarial loss from $\mathbf{D_1}$ to $\mathbf{G}$ on the target prediction, which would encourage $\mathbf{G}$ to produce more similar segmentation distributions in the target domain to the source prediction. The discriminator loss and adversarial loss can be defined as the following:

$$\mathcal{L}_{adv_1}(\mathbf{I}_t) = \mathbb{E}_{x_t \sim \mathbf{I}_t} \log(\mathbf{1} - \mathbf{D}_1(\mathbf{S}_t)), \tag{1}$$

$$\begin{aligned} \mathcal{L}_{D_1}(\mathbf{I}_s, \mathbf{I}_t) = \mathbb{E}_{x_t \sim \mathbf{I}_t} \log(\mathbf{D}_1(\mathbf{S}_t)) + \\ \mathbb{E}_{x_s \sim \mathbf{I}_s} \log(\mathbf{1} - \mathbf{D}_1(\mathbf{S}_s)), \end{aligned} \tag{2}$$

Where $\mathbf{S}_s$ and $\mathbf{S}_t$ are the segmentation output from $\mathbb{D}_s$ and $\mathbb{D}_t$, respectively.

C. Entropy Minimisation

Entropy minimisation showed great performance in the computer vision field for semantic segmentation. It is a regularisation method, which can prevent the model from overfitting and improve generalisation performance and robustness [20,28]. Entropy allows the exploration of the structural consistency between the two domains. For example, the cardiac LGE and cardiac CT images have no clear boundaries around the left ventricle (LV) and myocardium (Myo) , which makes it more difficult to segment in comparison to the bSSFP-MRI. Therefore, it enforces the segmentation network to generate high-entropy (uncertainty) on the soft boundary regions. To tackle the issue of uncertain predictions, we further adopt an entropy-driven adversarial learning model similar to [20] to reduce the performance gap between the source and target domains by enforcing the entropy maps of the target domain predictions to be similar to the source ones. Given the pixel-wise mask probability prediction S_s of input image $\mathbf{I}_s$, we use the Shannon Entropy to calculate the entropy map at pixel-level [20] shown in Eq. 3:

$$\mathbf{E}_s(\mathbf{I}_s) = -\frac{1}{N} \sum_{n=1}^{N} \sum_{c=1}^{C} \mathbf{S}_s^{n,c} log(\mathbf{S}_s^{n,c}) \tag{3}$$

where n indicates image number and c indicates channel number. To conduct the entropy-driven adversarial learning, we constructed $\mathbf{D}_2$ as an entropy discriminator network to align the entropy maps between $\mathbf{E}_s$ and $\mathbf{E}_t$ and enforce the segmentation network to minimise the entropy in the target domain. Similar to output space alignment, the entropy discriminator aims to figure out whether the entropy map is

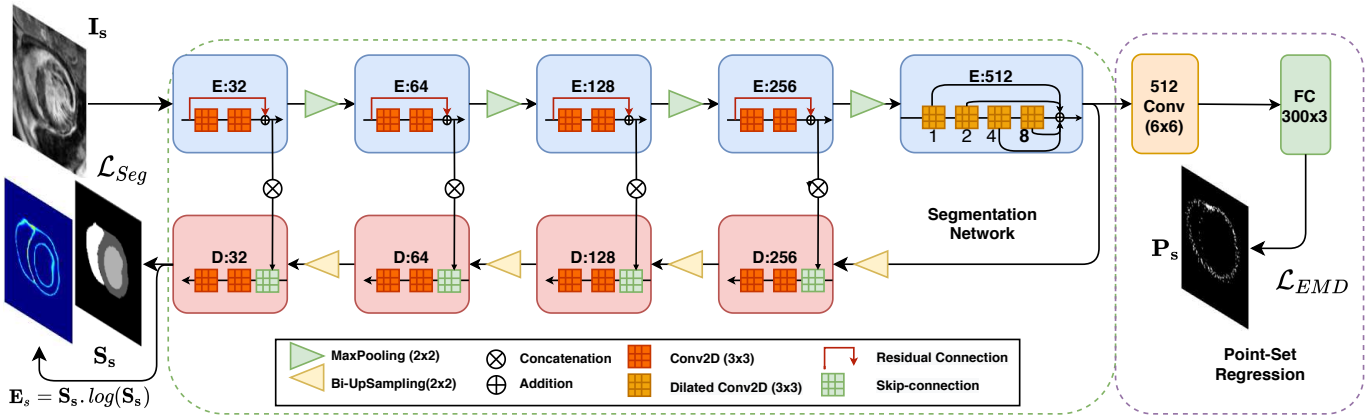


Fig. 2. Multi-task segmentation and point-cloud regression network architecture. An input image from the source or target domain is fed into the model and the model predicts three outputs including, a segmentation probability map $\mathbf{S}_s$, an output entropy map $\mathbf{E}_s$ and a point-cloud vector $\mathbf{P}_s$, respectively.

from the source or the target domain.

$$\mathcal{L}_{adv_2}(\mathbf{I}_t) = \mathbb{E}_{x_t \sim \mathbf{I}_t} \log(\mathbf{1} - \mathbf{D}_2(\mathbf{E}_t)), \tag{4}$$

$$\begin{aligned}\mathcal{L}_{D_2}(\mathbf{I}_s, \mathbf{I}_t) = \mathbb{E}_{x_t \sim \mathbf{I}_t} \log(\mathbf{D}_2(\mathbf{E}_t)) + \\ \mathbb{E}_{x_s \sim \mathbf{I}_s} \log(\mathbf{1} - \mathbf{D}_2(\mathbf{E}_s)),\end{aligned} \tag{5}$$

Where $\mathbf{E}_s$ and $\mathbf{E}_t$ are the self-weighted information maps from $\mathbb{D}_s$ and $\mathbb{D}_t$, respectively (cf. Eq. 3).

D. Point-cloud Shape Alignment

Incorporating shape information could improve the performance of a medical image segmentation model. One way to generate an overall shape of an anatomical object is using "point-clouds". Point-clouds contain data that expresses the external surface of an object onto a three-dimensional structure using the euclidean x, y, z coordinates. It is represented as a set of 3D points $P_i | i = 1, ..., n$, where each point P_i is a vector of its (x, y, z) coordinates.

Learning shape representation with point-clouds for target structures with distinct anatomy could be useful for the segmentation of target shapes. Hence, we developed a novel UDA using point-clouds shape learning combined with feature adaptation to improve the segmentation performance of our UDA. Our segmentation network has another head attached that takes deep encoded features containing object shape information in the form of point-clouds. Our model outputs shape representations of target cardiac structure; where the shape representations are sets of points located on the combined surface of LV, RV, and Myo. Our point-cloud network is inspired by point set generator (Fan *et al.* [36]). Here the authors use a point-cloud generator which is built with 2D CNN layers. It takes 2D images and random vectors as its inputs and outputs sets of points as the 3D reconstruction of target objects. To obtain 3D reconstruction, a moderate level of uncertainty is desirable and useful, which Fan *et al.* achieve by the use of the random vectors [36].

In our multi-task learning setup, the proposed point-cloud regressor is jointly optimised, with the DR-UNet segmentation model. Fig. 2 shows the pipeline of our multi-task learning network. This model directly consumes unordered point-clouds as inputs. The point-cloud regressor network takes the deep latent features extracted by DR-UNet encoder. It has a convolution layer with a kernel size 6×6 and a fully connected layer which predicts a vector of 300×3 that represent the point-clouds of the target image.

1) Point-cloud Ground-Truth Generation: To generate ground-truth point-clouds, we combined LV, RV, and Myo annotations to produce the external heart surface as a binary mask. The ground-truth surface points are generated using the marching cubes algorithm and farthest point sampling [36]. The size of the point-cloud (N_P) is empirically set to 300 in all experiments, with coordinates normalised in the range of $(0, 1)$. The regression loss is measured with Earth Mover's Distance (EMD) metric [36].

2) Point-cloud Objective Function: To compare the predicted point-cloud matrix and the ground-truth, we need a suitable loss function that includes the inherent uncertainty of the point-cloud predictions. There are several loss functions for point-cloud estimation such as the Hausdorff distance (HD) and Chamfer distance (CD) [36]. However, these objective functions are not robust to the outliers. Here, to train our point-cloud regression network, we employed Earth Mover's Distance (EMD) [36]. Consider $\mathbf{P}_s$, $\mathbf{P}_{gt} \in \mathbb{R}^3$ of equal size $|\mathbf{P}_s| = |\mathbf{P}_{gt}|$, where $\mathbf{P}_s$ is the prediction and $\mathbf{P}_{gt}$ is the ground truth. The EMD loss between $\mathbf{P}_s$ and $\mathbf{P}_{gt}$ is defined as:

$$\mathcal{L}_{EMD}(\mathbf{P}_s, \mathbf{P}_{gt}) = \underset{\theta:\mathbf{P}_s \to \mathbf{P}_{gt}}{\arg\min} \Sigma_{x \in \mathbf{P}_s} ||x - \theta(x)||_2 \tag{6}$$

where $\theta : \mathbf{P}_s \to \mathbf{P}_{gt}$ is a one-to-one correspondence. The EMD tries to solve an optimisation problem called the assignment problem. For every but a zero measure subset of point set pairs, the optimal one-to-one correspondence θ is unique and invariant under the infinite flow of the points. Therefore, EMD is differentiable practically everywhere.

E. Point-Net Discriminator

The main problem with point-clouds is that common convolutional architecture expects highly regular input data format,

like the image or temporal features. To be able to adapt the point-cloud features between two domains, there is a need for a network which solely processes point-clouds. Inspired by "PointNet" architecture [37], we propose a discriminator for adversarial adaptation of source and target domains based on point-clouds. This network learns a spatial encoding of every point within the point-clouds and then aggregate all the features to a global vector. Fig. 3. presents the overall architecture of our proposed PointNet discriminator network $\mathbf{D_3}$. The discriminator network receives 300×3 points as input and discriminates whether it is from source or target domain. In this network, we employ transformation layers proposed in [37] to enable learned representation by the network to be invariant to geometric alterations. For all the layers, we used Batch Normalisation with the ReLU activation function. There are fully connected layers in the network that aggregate learned optimal values into the global descriptor for the complete shape. To train the discriminator binary cross-entropy loss is used for adversarial learning, which identifies the generated points $\mathbf{P}_s$ for source domain versus target domain $\mathbf{P}_t$. The adversarial loss and discriminator loss is similar to entropy and output space alignment, which is defined as follows:

$$\mathcal{L}_{adv_3}(\mathbf{I}_t) = \mathbb{E}_{x_t \sim \mathbf{I}_t} \log(\mathbf{1} - \mathbf{D}_3(\mathbf{P}_t)), \tag{7}$$

$$\begin{aligned} \mathcal{L}_{D_3}(\mathbf{I}_s, \mathbf{I}_t) = \mathbb{E}_{x_t \sim \mathbf{I}_t} \log(\mathbf{D}_3(\mathbf{P}_t)) + \\ \mathbb{E}_{x_s \sim \mathbf{I}_s} \log(\mathbf{1} - \mathbf{D}_3(\mathbf{P}_s)), \end{aligned} \tag{8}$$

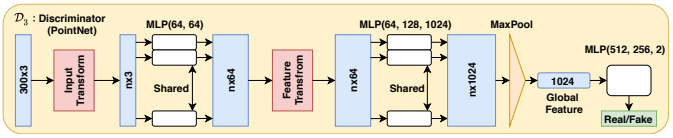


Fig. 3. PointNet discriminator architecture. It takes an input size of $\mathbf{300 \times 3}$ and applies a geometric transformation on input and feature level. The MaxPooling layer shrinks the features to generate global features, and there are three fully connected layers with a *softmax* activation function that classify the point-cloud as real(source) or fake(target).

Total Objective Function: To train the proposed method in an end-to-end fashion, we use the binary cross-entropy + Jaccard loss ($\mathcal{L}_{seg}$) [38] as the segmentation loss for supervised training of source domain. To obtain the total loss, we combine Eq. 1, 2, 4, 5, 6, 7 and 8 which can be formulated as below.

$$\begin{aligned} \mathcal{L}_{total}(\mathbf{I}_s, \mathbf{I}_t) = \underbrace{\mathcal{L}_{seg}(\mathbf{I}_s) + \mathcal{L}_{EMD}(\mathbf{I}_s)}_{\text{supervised}} + \\ \underbrace{\lambda_{adv_1}\mathcal{L}_{adv1}(\mathbf{I}_t) - \lambda_{D_1}\mathcal{L}_{D_1}(\mathbf{I}_s, \mathbf{I}_t)}_{\text{output space}} + \\ \underbrace{\lambda_{adv_2}\mathcal{L}_{adv2}(\mathbf{I}_t) - \lambda_{D_2}\mathcal{L}_{D_2}(\mathbf{I}_s, \mathbf{I}_t)}_{\text{entropy space}} + \\ \underbrace{\lambda_{adv_3}\mathcal{L}_{adv3}(\mathbf{I}_t) - \lambda_{D_3}\mathcal{L}_{D_3}(\mathbf{I}_s, \mathbf{I}_t)}_{\text{Point-cloud space}} \end{aligned} \tag{9}$$

where $\lambda_{adv_1}, \lambda_{D_1}, \lambda_{adv_2}, \lambda_{D_2}, \lambda_{adv_3}, \lambda_{D_3}$ are weights to balance the losses. We empirically set these values to 1 for adversarial losses and 0.2 for discriminator losses. The overall min-max optimisation problem can be summarised as following:

$$\arg\min_{\mathbb{D}_s, \mathbb{D}_t} \arg\max_{\mathbf{D}_1, \mathbf{D}_2, \mathbf{D}_3} \mathcal{L}_{total}(\mathbf{I}_s, \mathbf{I}_t) \tag{10}$$

where $\mathbb{D}_s$ and $\mathbb{D}_t$ are networks with shared trainable weights.

IV. Experimental Results

A. Datasets

To validate our proposed method for UDA segmentation, we utilised two datasets.

1) Cardiac Multi-sequence Segmentation.: This STACOM MS-CMRSeg [12] 2019 challenge dataset contains 45 short-axis Cine-MRI scans from patients diagnosed with cardiomyopathy. The dataset was collected in Shanghai Renji hospital and institutional ethics approved for all patients data [5]. For each patient, three different Cine-MR sequences were acquired, *viz.* bSSFP, LGE, and T2-weighted. The ground-truth contours were generated by two experts and include the right ventricle (RV) cavity, the LV cavity, and the myocardium region. The LGE-MRI is a T1-weighted sequence with inversion-recovery, consisting of $10-18$ slices comprising the main section of the ventricles. These images have a resolution of 512×512 pixels, with an in-plane resolution of 0.75×0.75 mm and a slice thickness of 5 mm. The bSSFP-MRI sequences consist of $8 - 12$ slices, including the full ventricles from the basal plane of the mitral valve to the apex. These have an image resolution of 1.25×1.25 mm and a slice thickness of $8 - 13$ mm. The T2-weighted-MRI sequences have fewer number of slices between $3 - 7$, with an in-plane resolution of 1.35×1.35 mm and slice thickness is $12 - 20$ mm. Since T2-weighted and bSSFP sequences had very few slices, we combined both (henceforth called bSSFP-CMR) as the source domain and used LGE-MRI as the target domain. The images within MS-CMRSeg 2019 dataset has a high level of variation in contrast and brightness. The variability results from different system settings and data acquisition, which makes it harder for neural networks to process the images. Therefore, we enhanced the cine-MR image contrast slice wise, using histogram equalisation. The MR sequences are normalised using min-max normalisation and centre-cropped to 224×224 pixels to have only region-of-interest (ROIs) areas. However, we found this is not an optimal solution because cardiac anatomy can exist in different regions of MRI scan. Therefore, we have also implemented an ROI detector. The ROI detector is a shallow UNet [39] that takes the binary mask of original images and predicts a coarse segmentation of RVs and LVs. Then we find the centre of the predicted segmentation mask and crop the real images with a size of 224×224 pixels based on this reference.

2) Cardiac Cross-modality Segmentation.: To further evaluate our proposed method, we employed the Multi-Modality Whole Heart Segmentation (MM-WHS) Challenge 2017 dataset for cardiac segmentation [13]. This dataset contains 20 MRI and 20 CT unpaired volumes with ground truth masks. For evaluating the domain adaptation on cardiac segmentation, we include the following four structures: ascending aorta (AA), the left atrium blood cavity (LA), the LV blood cavity,

and the myocardium of the LV. We used the same data split as in [27] for training (16 subjects) and testing (4 subjects) subsets in experiments. The dataset is prepossessed such that the MRI and CT images are re-oriented, resized and cropped centring at the heart region, such that the view of multi-modal images are roughly on the same page. We extract MRI and CT scans by 2D slices of size 256×256 at coronal plane during training and the whole dataset is normalised in 3D for each modality respectively. Data augmentations like rotation, zoom and affine transformation are utilised during the training.

3) Evaluation Metrics: We employed commonly-used metrics, the Dice similarity coefficient (Dice), the Hausdorff distance (HD) [12] and Average surface distance (ASD) [27], to quantitatively evaluate the segmentation performance of models. Dice measures the voxel-wise segmentation accuracy between the predicted and reference volumes. HD and ASD calculate the maximum and average distances between the surface of the prediction mask and the ground-truth in 3D. Hence, a higher Dice value and a lower HD and ASD values indicate better segmentation. The evaluation is performed on the subject-level segmentation volume, to be consistent with the MS-CMRSeg challenge benchmark study as well as previous works [10,27,30,34,40].

4) Implementation details: We implemented our framework with the PyTorch 1.4 deep learning library. We trained the whole pipeline directly without any warm-up phase of supervised learning with a mini-batch of size 16. The discriminators $\mathbf{D}_1$, $\mathbf{D}_2$, and $\mathbf{D}_3$, were optimised with the SGD algorithm, while the Adam optimiser is utilised for the segmentation network $\mathbf{G}$. For MS-CMRSeg dataset, we set the initial learning rate of Adam as 0.001 and reduced it by a factor of 0.2 every 100 epochs for a total of 600 epochs. The learning rate of discriminator training was set to 0.000025.

For MM-WHS dataset, the learning rate for the segmentation network $\mathbf{G}$ is set to 0.0002 without any learning rate decay. Compared with the learning rate setting of the MS-CMRSeg dataset, we have more meticulous configuration for the discriminators and the corresponding adversarial learning for the MM-WHS dataset. The models were trained from scratch until there is no further improvement for the last 100 epochs. We employed the same preprocessing and normalisation scheme as in [27]. For a fair comparison with other methods, we do not apply any extra data augmentations.

B. Comparison with other methods

To demonstrate the effectiveness of our proposed UDA method for leveraging multi-modal data, we compare it with both supervised and unsupervised learning methods.

MS-CMRSeg Dataset: Table I shows the quantitative results of different algorithms for the MS-CMRSeg challenge, specifically for the 40 LGE-MRI segmentation test data. We first compare with the model trained with only limited labelled LGE-MRI data $\mathbb{D}_t$ (referred as Supervised only) and take all other methods (Unsupervised, Inter-Ob) for benchmark study including ADVENT [20] and AdaptSeg [19] for comparison. Besides two-stage approaches, we also compare with the end-to-end model and inter-observer study. Meanwhile, we also compare with segmentation methods like Chen *et al.* [10] (Unsupervised) and Wang *et al.* [42] (Supervised), that achieved SOTA performance in MS-CMRSeg [12] cardiac challenge segmentation. Supervised methods in MS-CMRSeg challenge had access to only 5 LGE-MRI studies to train their models where 2 or 3 studies are used for training and cross-validation.

As shown in Table I, the baseline method without DA performed poorly, which emphasises the domain shift between the source (bSSFP-MRI) and the target domain (LGE-MRI). The best model with supervised training regime achieved 88.6% mean Dice and 11.57 HD score by taking only limited labelled target data (5 LGE-MRI subjects) during training. When two types of data sources are available, unsupervised methods achieve comparable Dice scores to the supervised-only model by utilising adversarial training or generating synthetic data. Compared with the supervised-only methods, Chen *et al.* [10] approach further improved the segmentation performance, demonstrating the effectiveness of leveraging multi-sequence data $\mathbb{D}_s$ and unlabelled data $\mathbb{D}_t$ for training.

The volumetric Dice scores of our proposed method are 0.794 ± 0.041 (Myo), 0.909 ± 0.032 (LV), and 0.878 ± 0.053 (RV), respectively and the volumetric HD values are respectively $9.390 \pm 2.628mm$ (Myo), $7.647 \pm 3.231mm$ (LV), and $8.452 \pm 2.831mm$ (RV). The average Dice score of our method ranks second in the Table I, and it is 4 points higher as compared to the inter-observer Dice scores. Our method achieved the lowest average HD score in comparison to all existing approaches. The main reason can be in two folds. First, our network employs point-cloud to incorporate shape information during training which enforces explicitly in producing a more smooth surface and accurate segmentation output. Secondly, the concurrent adaptation in different spaces could lead to a better model optimisation. Interestingly, the unsupervised methods performed comparably to supervised ones. Specifically, LV structure has a different geometry shape from base to apex, but it is simpler to segment and has higher Dice scores in comparison to RV and Myo in most of the methods. We reported the slice-wise accuracy of different methods for different positions in the discussion section (Table V). The myocardial wall is typically the challenging structure to segment, especially in this dataset, in which some of the cases in the test data included scars. In comparison to the inter-observer study, our method achieved higher Dice scores for RV and Myo with 6.0% and 3.0% improvements, respectively. Fig. 4 illustrates some qualitative examples comparing our proposed method with other SOTA approaches along with the baseline. Our model shows better results on LV classes, while ADVENT sometimes makes mistakes of predicting myocardial wall (1^{st} row, 5^{th} column). It is observed that our method better identifies heart substructures with a clean and accurate boundary, and produces less false positive predictions and more similar results to the ground-truth (9^{th} & 10^{th} columns) compared with other methods.

MM-WHS Dataset: Table II shows the quantitative results of different algorithms for the MM-WHS challenge dataset, specifically for the 4 CT segmentation test data. Similar to the previous dataset, we have compared our model with the baseline model without domain adaptation, feature-based and

TABLE I

PERFORMANCE COMPARISON BETWEEN OUR PROPOSED METHOD AND DIFFERENT UNSUPERVISED AND SUPERVISED DOMAIN ADAPTION METHODS FOR CARDIAC MULTI-MODAL SEGMENTATION BASED ON VOLUMETRIC DICE AND HD. THE VALUES ARE SHOWN WITH $(mean \pm std)$.

Methods	Volumetric Dice $[mean \pm std] \uparrow$				Volumetric HD [mm] $\downarrow$				Training
	Myo	LV	RV	Average Dice	Myo	LV	RV	Average HD	
Baseline (W/o DA)	0.392±0.210	0.651±0.265	0.619±0.270	0.554±0.248	20.54±7.582	15.88±10.723	22.32±17.999	19.59±12.101	Unsupervised
Chen et al. [10]	**0.826±0.035**	**0.919±0.026**	0.875±0.050	**0.873±0.037**	10.28±3.376	12.45±3.142	15.38±6.942	12.703±4.487	Unsupervised
Proposed method	0.794±0.041	0.909±0.032	**0.878±0.053**	0.860±0.042	**9.390±2.628**	**7.647±3.231**	**8.452±2.831**	**8.496±2.897**	**Unsupervised**
ADVENT [20]	0.778±0.061	0.906±0.034	0.867±0.063	0.850±0.053	9.727±2.686	7.375±3.311	9.952±3.459	9.018±3.152	Unsupervised
Wang et al. [31]	0.796±0.059	0.896±0.047	0.846±0.086	0.846±0.064	13.59±5.206	15.7±5.814	15.21±6.327	14.833±5.782	Unsupervised
Ly et al. [41]	0.705±0.115	0.87±0.051	0.762±0.150	0.779±0.105	41.74±7.696	42.79±13.26	34.38±8.065	39.637±9.674	Unsupervised
Wang et al. [42]	0.843±0.048	0.926±0.028	0.890±0.044	0.886±0.04	9.748±3.28	11.65±4.002	13.34±4.615	11.579±3.966	Supervised
Campello et al. [43]	0.81±0.061	0.898±0.045	0.866±0.050	0.858±0.052	10.78±4.066	11.96±3.62	15.91±6.895	12.883±4.86	Supervised
Vesal et al. [34]	0.789±0.073	0.912±0.034	0.833±0.084	0.845±0.064	11.29±4.559	12.54±3.379	17.11±6.141	13.647±4.693	Supervised
Roth et al. [32]	0.78±0.047	0.89±0.043	0.844±0.063	0.838±0.051	11.58±7.524	16.25±6.336	18.12±9.262	15.317±7.707	Supervised
Liu et al. [44]	0.751±0.119	0.884±0.07	0.791±0.165	0.809±0.118	14.30±8.17	14.75±7.823	17.87±9.322	15.64±8.438	Supervised
Chen et al. [45]	0.61±0.102	0.824±0.068	0.71±0.135	0.715±0.102	23.69±14.66	24.62±12.66	23.46±7.596	23.923±11.639	Supervised
Inter-Observer [12]	0.764±0.069	0.881±0.064	0.816±0.084	0.82±0.072	12.03±4.443	14.32±5.164	21.53±9.46	15.96±6.356	Inter-Ob

TABLE II

PERFORMANCE COMPARISON BETWEEN OUR PROPOSED METHOD AND DIFFERENT UDA METHODS FOR CARDIAC CROSS-MODALITY SEGMENTATION (MRI $\rightarrow$ CT) BASED ON VOLUMETRIC DICE AND ASD. THE VALUES ARE SHOWN WITH $(mean)$.

Methods	Volumetric Dice $[mean \pm std] \uparrow$					Volumetric ASD [mm] $\downarrow$				
	AA	LA	LV	Myo	Average Dice	AA	LA	LV	Myo	Average HD
Baseline (W/o DA)	0.303	0.846	0.360	0.517	0.507	18.74	2.30	13.10	7.43	10.39
Baseline (W/o DA) + Point-Cloud	0.787	0.744	0.315	0.335	0.545	3.94	2.94	43.18	12.00	15.51
CycleGAN [14]	0.738	0.757	0.523	0.287	0.576	11.50	13.60	9.20	8.80	10.80
ADVENT [20]	0.812	0.765	0.329	0.225	0.533	3.68	3.63	15.41	28.71	12.86
PnP-AdaNet [27]	0.740	0.689	0.619	0.508	0.639	12.80	6.30	17.40	14.70	12.80
SIFA [7]	0.811	0.764	**0.757**	**0.587**	**0.730**	10.60	7.40	6.70	7.80	8.10
Proposed Method	**0.830**	**0.813**	0.672	0.584	0.725	**2.90**	**2.66**	**6.32**	**6.38**	**4.56**

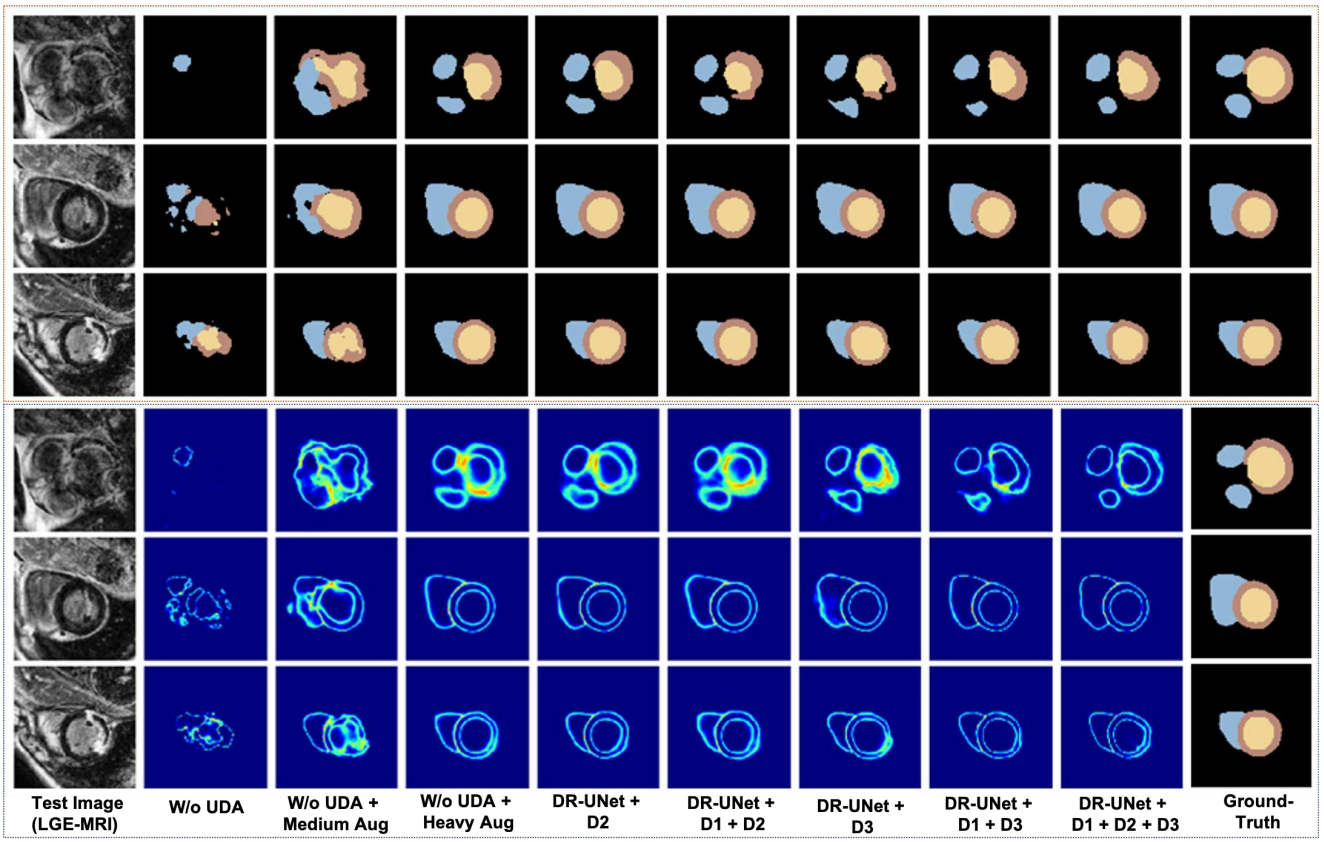


Fig. 4. A visual comparison of segmentation output and entropy map results produced by different methods for LGE images. From left to right are the raw LGE test images (1st column), W/o Adaptation (2nd column), W/o Adaptation but with medium and high-level augmentation (3rd & 4th column), results of our UDA segmentation methods (5th-9th column), and ground-truth (last column). The LV, Myo, and RV are indicated, in yellow, brown, and blue colour, respectively. Each row corresponds to one example.

image-based UDA segmentation methods, including CycleGAN [14], ADVENT [20,27] and SIFA [7] that achieved SOTA performance in MM-WHS [13] benchmark dataset. For this dataset, we considered ASD metric and not HD, since almost all the existing methods reported their model performance with ASD. As a baseline study, we obtained the "Baseline W/o DA" lower bound by directly applying the model learned in MR source domain to target CT images without using any domain adaptation method. The poor performance of the baseline model indicates that the domain shift between the source and the target domain is significant. The baseline method after adding the point-cloud head improved the average volumetric Dice score to 54.5%, but produced a very high average ASD of $15.51mm$.

Remarkably, with our proposed network, the average Dice score improved to 72.5%, and the ASD score reduced to $4.56mm$. We achieved $83.0\%, 81.3\%, 67.2\%$ Dice scores for AA structure, LA and LV blood pools respectively. Notably, compared to SIFA model, which conducts both image and feature adaptations, our method achieved superior performance especially for the AA and LA structures, which have limited contrast in CT images. However, it produced a slightly worse segmentation output for Myo. The proposed network outperforms CycleGAN, ADVENT and PnP-AdaNet for segmentation of all cardiac structures. Fig. 6 shows some qualitative examples comparing our proposed method with other SOTA approaches along with the baseline model for MM-WHS CT test images. As it can be seen in the 3rd row in Fig. 6, the LV and Myo structures have very limited intensity contrast with their surrounding tissues, but our method can make good predictions while all the other methods fail in this challenging case. Similar to MS-CMRSeg dataset, our model achieved the lowest average ASD score among all the networks. Hence, the point-cloud adaptation enforced the model to avoid over-segmentation of the cardiac structure and produced more smooth surface shape.

C. Ablation study

To evaluate whether adapting features in different spaces is truly beneficial for accurate cardiac segmentation on target modality, we performed an ablation study on the MS-CMRSeg benchmark dataset. First, we evaluate the effectiveness of data augmentation on the target domain. We test three different data augmentations strategies before implementing UDA. In the first attempt, we trained the segmentation network without any data augmentation, and the model achieved a mean volumetric Dice score of 0.554 ± 0.248 and an HD value of $19.584 \pm 12.101mm$ (1st data row, Table III) on 40 LGE-MRI subjects of the target domain. In the next step, we applied histogram equalisation and rotate the myocardium with two angles, $[30°, 60°]$ [43], to increase the number of training samples. With this augmentation, we observed a 16.0% improvement in mean volumetric Dice and minor improvement in volumetric HD (2nd data row, Table III). The reason for high volumetric HD values is mainly because the model could not learn the noisy border of LGE-images, which also has low contrast in those regions. Ultimately, we utilised an online imaging library (imgAug [46]) and employed a set of extreme and complex augmentation such as affine-transformation, translation, gaussian noise, elastic deformation, contrast enhancement, optical deformation, etc. The segmentation network under these settings achieved a volumetric Dice score of 0.833 ± 0.063 and volumetric HD drastically reduced to $10.027 \pm 3.985mm$ (3rd data row, Table III).

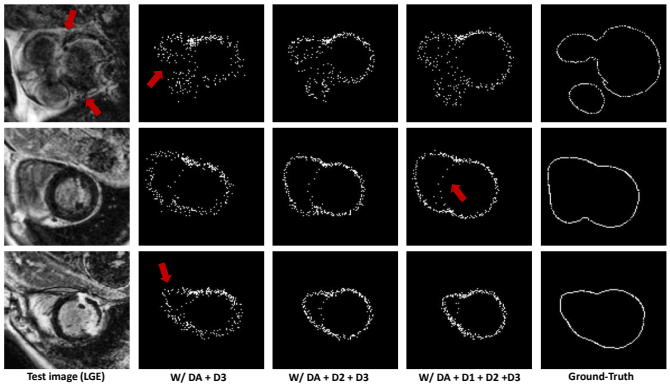


Fig. 5. Visual comparison of point-clouds regression results produced by different methods for test LGE-MR images. From left to right are the raw LGE test images (1st column) and the results of our UDA methods (2nd-4th column), and ground-truth point-cloud (last column). The red arrows point at the regions reconstructed by the model because of the noise in the image.

Furthermore, we demonstrate that minimising the entropy in the target domain and adapting the shape using point-clouds can work jointly to improve domain adaptation performance. The quantitative and visual experimental analysis are shown in Table III and Fig. 4 respectively. Our baseline network uses entropy minimisation only, which is constructed by removing the output space alignment and point-cloud adversarial loss when training the network. Compared with the "W/o UDA" lower bound, our baseline network with image alignment (D2) alone (4th data row, Table III) increased the average Dice to 85.0% (1.7% improvement compared to heavy augmentation). It shows that with reducing uncertain regions via entropy maps in the target domain, the target images have been brought closer to the source domain successfully. Then we add the output space alignment (D1) in the semantic prediction space, which slightly improved the average Dice (1.7% improvement) and also achieved a good performance jump for the HD value (5th data row, Table III). Finally, by adding the point-could shape alignment network, the model was able to obtain an average Dice of 86.0% (8th data row, Table III), which, if compared with W/o UDA + augmentation, improved the segmentation performance by 3.3%. The incremental increase in segmentation accuracy explains that feature adaptation and incorporating shape prior can be jointly conducted to achieve better domain adaptation. Feature alignment in different compact spaces could inject effects from integral aspects to encourage domain invariance. The last row of the Table III shows the Dice and HD values, when we train our model in a multi-task fashion without any discriminators and adversarial learning.

A very similar observation was achieved for the second benchmark dataset (MRI $\rightarrow$ CT). However, it is noticeable

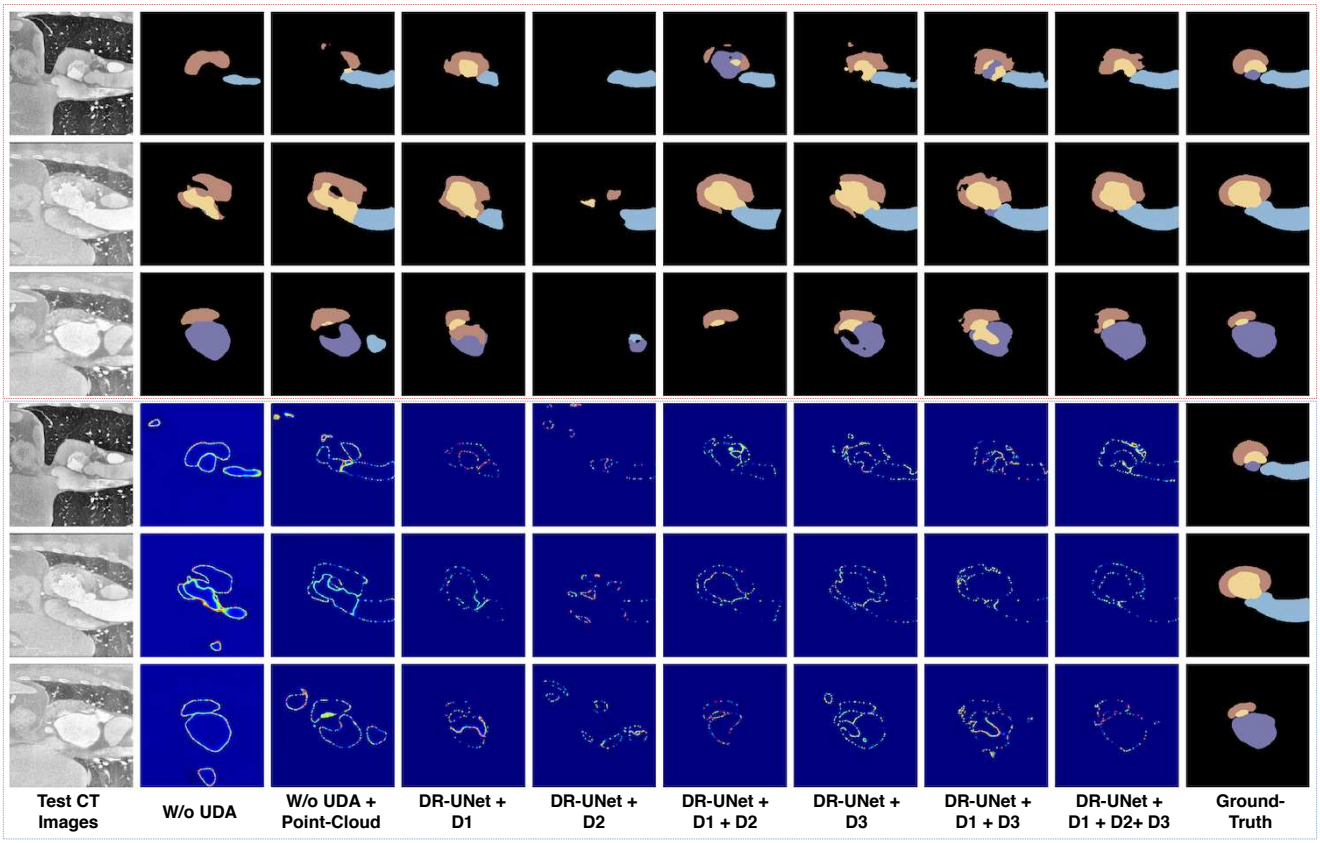


Fig. 6. A visual comparison of segmentation output and entropy map results produced by different methods for CT images. From left to right are the raw CT test images (1st column), W/o Adaptation (2nd column), W/o Adaptation with point-cloud as a multi-task(3rd column), results of our UDA segmentation methods (4th-9th column), and ground-truth (last column). The cardiac structures of AA, LA, LV, and Myo are indicated in blue, purple, yellow, and brown color respectively.

that the anatomy variation between MRI and CT modalities is quite significant, and it is a more challenging task compared to multi-sequence UDA. Table IV summarised the ablation study for MM-WHS dataset. In this experiment, the model with only entropy minimisation (D2) achieved an average volumetric Dice of 53.3% and ASD volumetric value of $12.86mm$ (3rd data row, Table IV), which is considerably lower than the model trained with output space alignment (D1) with an average Dice of 65.6% and ASD of $5.97mm$ (2nd data row, Table IV). Furthermore, when the model trained by employing both $\mathbf{D}_1 + \mathbf{D}_2$ (4th data row, Table IV) discriminators outperformed the individual models. It proves our initial hypothesis that enforcing entropy minimisation as an extra layer of adaptation can lead to a better model generalisation. Additionally, the model with point-cloud adaption (D3, 5th data row) without any other discriminators achieved comparable performance to $\mathbf{D}_1 + \mathbf{D}_2$ model. By comparing these two results, we can conclude that shape adaptation alone has a significant role for UDA when there is a high variation between source and target domains. Finally, the model with $\mathbf{D}_1 + \mathbf{D}_2 + \mathbf{D}_3$ outperformed the rest and achieved the best Dice and ASD scores across all cardiac structures. This model achieved an average volumetric Dice score of 72.5% and an ASD score of $4.56mm$, and improved the segmentation accuracy drastically, with a 22% jump compared to W/o UDA method.

The results for the ablation study (MM-WHS dataset) highlights the contribution of adapting features in different spaces by our proposed model. There is an incremental improvement in terms of evaluation metrics after adding every discriminator. On the other hand, our proposed method performs only feature adaptation compared to other models such as SIFA (and CycleGAN) that applied both image and feature adaptation concurrently. Moreover, SIFA and CycleGAN methods translate the appearance of the source domain images to the target domain for UDA, while our model does not need this step and therefore has less trainable parameters. Image-based UDA methods highly depend on the quality of the image translation and require more data to train the whole network, which is always not accessible. Hence, methods such as the proposed one can be an alternative solution when there is a lack of sufficient training samples.

1) Statistical Analysis: To statistically evaluate the correlation and agreement between our proposed UDA and ground-truth segmentation generated by experts, linear regression and Bland-Altman plots are used. The linear regression plots are linear fits to the scatter plot of the true values vs. the predicted data for LV volume. The linear regression plot (1st row, Fig. 7) showed that Pearson's correlation is superior ($r^2 = 0.95$) for our proposed approach, *DR-UNet + D1 + D2 + D3*, which includes entropy minimisation, point-cloud shape adaptation and output space alignment ($P < 0.01$ for all analyses, except W/o adaptation which achieved P-value of 0.11). These results show that our method has substantial correlations with ground-

TABLE III

ABLATION STUDY. EFFECTIVENESS OF DIFFERENT ADVERSARIAL DOMAIN ADAPTATION COMPONENTS IN OUR PROPOSED FRAMEWORK FOR MS-CMRSEG DATASET.

Cardiac bSSFP-MRI → Cardiac LGE-MRI												
Methods	Adversarial elements				Volumetric Dice [*mean* ± *std*] ↑				Volumetric HD [mm] ↓			
	$\mathbf{D_1}$	$\mathbf{D_2}$	$\mathbf{D_3}$	Aug	Myo	LV	RV	Average Dice	Myo	LV	RV	Average HD
W/o UDA	×	×	×	×	0.392±0.210	0.651±0.265	0.619±0.270	0.554±0.248	20.540±7.582	15.880±10.723	22.320±17.99	19.584±12.101
	×	×	×	✓	0.577±0.161	0.797±0.141	0.755±0.126	0.710±0.143	20.950±10.289	19.279±8.858	15.946±7.855	18.728±9.001
	×	×	×	✓	0.735±0.086	0.900±0.037	0.863±0.067	0.833±0.063	12.378±4.319	8.123±3.524	9.580±0.383	10.027±3.895
W/ UDA	×	✓	×	✓	0.778±0.061	0.906±0.034	0.867±0.063	0.850±0.053	9.727±2.686	7.375±3.311	9.952±3.459	9.018±3.152
	✓	✓	×	✓	0.776±0.07	0.904±0.04	0.869±0.067	0.85±0.059	11.483±4.538	8.005±2.977	9.509±3.946	9.666±3.82
	×	×	✓	✓	0.769±0.046	0.906±0.035	0.866±0.067	0.847±0.049	11.258±3.215	7.888±3.588	9.780±5.033	9.642±3.945
	×	✓	✓	✓	0.789±0.039	0.907±0.03	0.877±0.053	0.858±0.041	10.01±3.457	7.682±2.700	8.987±3.989	8.893±3.382
	✓	✓	✓	✓	**0.794±0.041**	**0.909±0.032**	**0.878±0.053**	**0.860±0.042**	**9.390±2.628**	**7.647±3.231**	**8.452±2.831**	**8.496±2.897**
Multi-Task	×	×	×	✓	0.78±0.049	0.908±0.028	0.872±0.053	0.853±0.043	10.140±3.136	7.710±2.908	10.007±3.649	9.286±3.231

TABLE IV

ABLATION STUDY. EFFECTIVENESS OF DIFFERENT ADVERSARIAL DOMAIN ADAPTATION COMPONENTS IN OUR PROPOSED FRAMEWORK FOR MM-WHS DATASET (MRI → CT).

Cardiac MRI → Cardiac CT														
Methods	Adversarial elements				Volumetric Dice $(mean)$ ↑					Volumentric ASD $[mm]$ ↓				
	D1	**D2**	**D3**	Aug	AA	LA	LV	Myo	Average	AA	LA	LV	Myo	Average
W/o UDA	×	×	×	×	0.303	0.846	0.360	0.517	0.507	18.74	2.30	13.10	7.43	10.39
W/ UDA	✓	×	×	✓	0.566	0.754	0.748	0.555	0.656	11.92	4.24	3.55	4.18	5.97
	×	✓	×	✓	0.812	0.765	0.329	0.225	0.533	3.68	3.63	15.41	28.71	12.86
	✓	✓	×	✓	0.794	0.782	0.583	0.571	0.683	4.42	3.37	8.64	5.44	5.47
	×	×	✓	✓	0.720	0.806	0.646	0.530	0.676	4.40	2.89	5.11	6.36	4.69
	×	✓	✓	✓	0.815	0.805	0.662	0.569	0.713	3.24	2.74	6.32	6.32	4.66
	✓	✓	✓	✓	**0.830**	**0.813**	**0.672**	**0.584**	**0.725**	**2.90**	**2.66**	**6.32**	**6.38**	**4.56**
Multi-task	×	×	×	×	0.787	0.744	0.315	0.335	0.545	3.94	2.94	43.18	12.00	15.51

truth segmentation by cardiologists. The Bland Altman plot shows limits of agreement (1.96 ∗ SD) at about 22 ml. These plots (2nd row, Fig. 7) demonstrated that the 95% limits of the agreement (14.25 mL) for our proposed method (*DR-UNet + D1 + D2 + D3*) are much higher, which asserts that our method is closer to the actual annotations. The similar observation has been achieved for MM-WHS test data, but due to space limitations, we have not included the plots.

V. DISCUSSION

For the diagnosis of cardiovascular diseases, several type of imaging modalities are used to measure the heart anatomy and morphology. Deep learning techniques demonstrated significant performance in obtaining accurate and reliable segmentation of multi-modal cardiac images. Nevertheless, these networks demand labelled data for each modality because deep models are not able to generalise well across modalities due to differences in data distributions. Data annotation problem was addressed with multi-modal image-to-image translation [11,14,24] to generate synthesised images and shift the appearance of the target images to the source domain. Recent studies investigated the application of cross-modality UDA to adapt deep features from the annotated source domain to the unlabelled target domain, achieving significant performances in computer vision [19,20] as well as in medical field [7,10,18,47]. With a similar theme, our shape-aware and entropy minimisation method manifests that, without further annotation, the UDA segmentation method can significantly diminish the domain shift and could achieve similar performance as semi/fully supervised methods.

In this paper, we present a novel interpretable UDA framework for multi-modal cardiac image segmentation. Our method integrates the entropy and point-clouds to leverage cross-modality priors from the source domain and to exploit the shape information embedded in the latent space of the segmentation network. Both adversarial learning components assist the model to learn invariant features for high entropy region and prior shape. Through multi-class segmentation tasks, we adequately demonstrated the overall effectiveness of our method without particular network architecture or hyperparameter tuning. It has been shown that a typical segmentation model like UNet or DR-UNet without domain adaptation has a high cross-modality performance degradation on the LGE-MRI and CT domains. However, our method achieved significant improvement in the segmentation for the cross-modality tasks.

The model (trained with an unsupervised learning strategy) that achieved the best Dice value of 87.3% (Table I) on the MS-MWSeg benchmark is not trained end-to-end. The authors generate stylised LGE-MRI images using MUNIT [22], thereby increasing the training data 10-fold. Then, they used a complex cascade UNet network architecture to refine their results. However, our method achieved a Dice value of 86.0%, using pure adversarial learning. The performance is not only similar to supervised and unsupervised models, but our model also outperformed all the methods by achieving the lowest HD value. The HD metric computes the smoothness of the object surface, and our point-clouds adaptation regressor could enforce the segmentation network in generating more smooth boundaries for ventricles. We further evaluated the accuracy of point-clouds regressor on the target domain (LGE-MRI) using EMD distance metric to see the impact of adversarial learning on three proposed methods (Fig. 8). *DR-UNet + D1 + D2 + D3* achieved the lowest EMD distance in comparison.

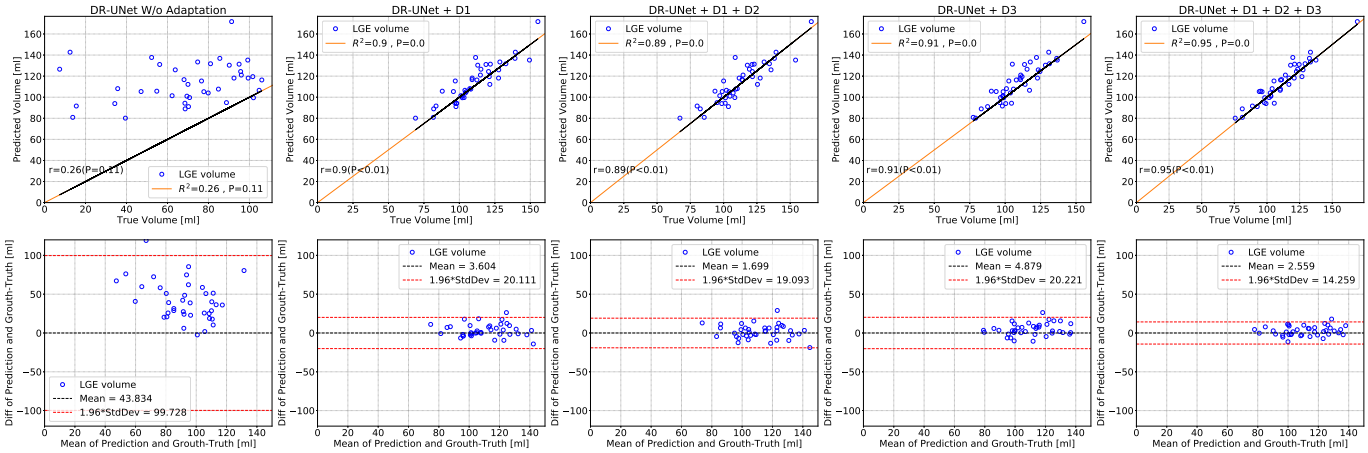


Fig. 7. Linear regression (top row) and Bland–Altman (bottom row) plots for the assessment of inter-observer variability for ventricular volume measurements by different methods. The volume measurements present an exemplary correlation ($\boldsymbol{R^2 = 0.95}$) with a P-value $< \mathbf{0.01}$ between our proposed UDA method (last column) and expert cardiologist annotation. Bland-Altman analysis plots show that the estimated ventricular volumes using our model is very close to the ground-truth (95% confidence intervals).

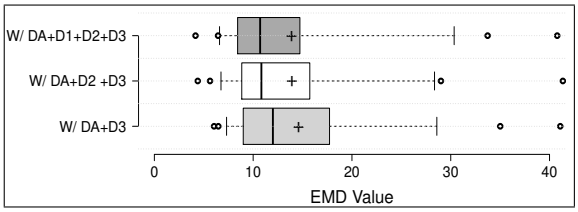


Fig. 8. Box-Plot of point-cloud EMD results in each ablation experimental setting for analysis of the proposed UDA framework.

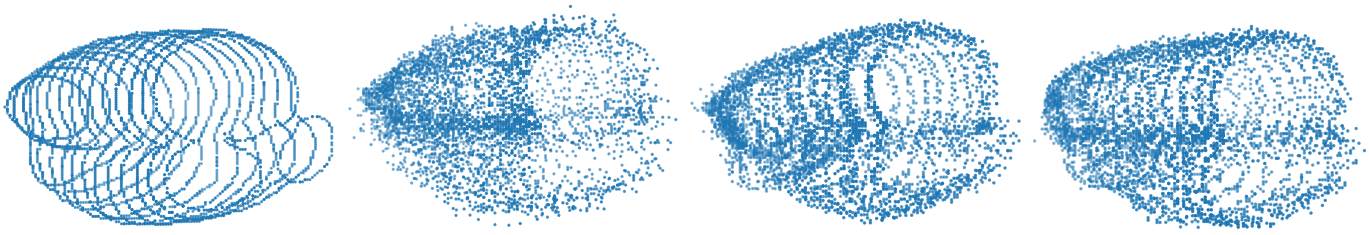
Furthermore, for qualitative evaluation, we visualised the 3D point-clouds by stacking the model prediction for a test LGE-MRI case (Fig. 9). A similar observation can be seen here as the generated point-clouds by *DR-UNet + D1 + D2 + D3* has lower distance error in comparison to the ground-truth, and the overall shape of ventricles is well-preserved.

Although the segmentation accuracy of our approach was high for the LV and RV blood pool, there is still room for improvement for the myocardium. In a few cases within the LGE-MRI dataset, the different myocardial wall and LV blood pool are obscure due to existing scars and edema in myocardium regions. A few such training cases makes the segmentation task more challenging. Nevertheless, our point-cloud classifier shows an interesting ability where it is able to detect the borders between left and right ventricles (2^{nd} row, 4^{th} column, Fig. 5). Improving the point-cloud estimation accuracy from binary to multi-class regression may help in such cases to show some distinction between the epicardium and endocardium wall, which could lead to an improvement in the segmentation network accuracy, respectively.

To have a more in-depth evaluation of the performance of our model, we computed the slice-wise segmentation accuracy from different positions of the ventricles, including the apical slices (Apex), mid-ventricular slices (Mid) and basal slices (Base) [12,48]. Table V demonstrates the slice-wise Dice and HD values that were computed, by averaging all the numbers from the Myo, LV, and RV cardiac structures. It can be observed that different methods achieved different performance regarding segmentation accuracy for the slices from different positions. The baseline (W/o) UDA showed the best performance on apex slices but achieved the lowest Dice score and highest HD for the middle and base slices.

Similar to MS-CMRSeg dataset, incremental improvement can also be observed on MM-WHS dataset by applying different discriminators for our proposed model. The boundaries between the cardiac structures in CT images are highly obscure compared to MR images making it hard to generalise the model trained on MR to CT images. Different from MS-MWSeg dataset, the difference in ablation study on cross-modality task is more significant. By comparing the baseline method W/o UDA and $\mathbf{D}_2$, we observe $\mathbf{D}_2$ generally performs better than W/o UDA, while $\mathbf{D}_2$ has higher variance. This poor performance can also be found in Fig. 6. The model with $\mathbf{D}_2$ has a completely wrong prediction on some of the substructures. By comparing the entropy rows, we find entropy is successfully reduced by applying $\mathbf{D}_2$. Although, a low entropy does not represent a better segmentation. Due to the large distribution discrepancy, prediction on the target domain should be poor. Minimising entropy of prediction with low certainty may encourage the model to generate wrong segmentations. However, the poor performance by applying entropy discriminator alone is addressed by utilising output-space and point-cloud discriminators, *i.e.*, column $\mathbf{D}_1 + \mathbf{D}_2$, $\mathbf{D}_2 + \mathbf{D}_3$ and $\mathbf{D}_1 + \mathbf{D}_2 + \mathbf{D}_3$.

In volumetric image segmentation, incorporating temporal information (all slices put together) could be useful. Since our entire pipeline is based on 2D CNN models, the temporal information (present in the form of sequence of slices) is not integrated while training. Nonetheless, our proposed method is able to use the temporal information in the form of point-clouds as prior information for adversarial learning. Our UDA segmentation has several components, including the adversarial training part, and implementing such a network under the same configuration in 3D (using temporal information) requires a high amount of memory and longer training time. Nevertheless, we believe that incorporating shape information using point-clouds in the 3D domain could lead to better



(a) (b) (c) (d)

Fig. 9. 3D visualisation of predicted point-clouds by our methods for LGE-MRI test data, after domain adaptation. (a) shows the ground-truth point-cloud, and (b-d) are the predictions for DR-UNet+D3, DR-UNet+D1+D2, and DR-UNet+D1+D2+D3, respectively.

TABLE V
SLICE-WISE ACCURACIES OF THE APICAL (APEX), MID-VENTRICULAR (MID), AND BASAL (BASE) SLICES FOR DIFFERENT UNSUPERVISED METHODS.

Methods	Slice-wise Dice [*mean* ± *std*] ↑				Slice-wise HD [mm] ↓			
	Apex	Mid	Base	Average	Apex	Mid	Base	Average
Baseline (W/o DA)	0.479±0.260	0.216±0.214	0.191±0.232	0.295±0.235	7.94±16.67	32.23±16.19	26.41±24.17	22.19±19.01
Chen et al. [10]	**0.706±0.203**	**0.883±0.078**	**0.857±0.140**	**0.815±0.140**	11.24±10.73	**6.140±5.14**	**8.22±7.30**	**8.53±7.72**
Proposed method	0.628±0.162	0.862±0.080	0.753±0.258	0.748±0.167	**5.22±4.57**	8.41±3.91	17.20±13.97	10.28±7.48
ADVENT [20]	0.623±0.237	0.859±0.086	0.746±0.205	0.743±0.176	5.49±3.68	9.10±5.02	20.47±15.27	11.69±7.99
Wang et al. [42]	0.502±0.292	0.844±0.136	0.828±0.165	0.725±0.198	20.76±17.64	7.604±8.77	8.99±7.75	12.45±11.38
Ly et al. [41]	0.612±0.230	0.795±0.142	0.724±0.221	0.710±0.198	14.45±15.52	12.74±13.57	16.99±15.93	14.72±15.00

performance improvement since the overall shape of cardiac anatomy will be taken into consideration in a temporal manner.

VI. CONCLUSION

In this paper, we proposed a novel entropy and shape-aware UDA method for multi-modal cardiac MR image segmentation. We approached the domain adaptation problem via adversarial learning in different spaces. We showed that introducing additional shape information using point-clouds along with entropy minimisation brings further complementary effects to bridge the performance gap between source and target domains. We construct a novel segmentation network such that the point-cloud information is embedded into a dedicated deep architecture using an auxiliary point-cloud regression task. A novel point-cloud discriminator based on PointNet is proposed to distinguish whether the point-clouds are from source or target domain. Experimental results on two benchmark cardiac image datasets highlighted that the proposed end-to-end method outperforms many baseline models and SOTA algorithms. For future work, we aim to extend our UDA segmentation method from 2D to 3D to process volumetric data for adaptation, which is clinically more relevant. To prove the generalisability and robustness of our method, we plan to test it on larger clinical cohort studies and employ it on other domains beyond cardiac segmentation task, such as multi-modal brain studies for lesion segmentation. We also aim to find a better way to choose the hyperparameters λ_1, λ_2 and λ_3 for the total loss function and learning rates, especially on how to efficiently control the effect of the gradient with respect to different discriminators for better model convergence.

REFERENCES

[1] H. W. Kim, A. Farzaneh-Far, and R. J. Kim, "Cardiovascular magnetic resonance in patients with myocardial infarction: Current and emerging applications," *Journal of the American College of Cardiology*, vol. 55, no. 1, pp. 1 – 16, 2009.

[2] E. S. D. Group *et al.*, "European Society of Cardiology: Cardiovascular Disease Statistics 2017," *European Heart Journal*, vol. 39, no. 7, pp. 508–579, 11 2017.

[3] X. Zhuang, "Multivariate mixture model for myocardial segmentation combining multi-source images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2019.

[4] S. Hammer-Hansen *et al.*, "Mechanisms for overestimating acute myocardial infarct size with gadolinium-enhanced cardiovascular magnetic resonance imaging in humans: a quantitative and kinetic study†," *European Heart Journal - Cardiovascular Imaging*, vol. 17, no. 1, pp. 76–84, 05 2015.

[5] X. Zhuang, "Multivariate mixture model for cardiac segmentation from multi-sequence mri," in *MICCAI*, 2016, pp. 581–588.

[6] W. Yan *et al.*, "The domain shift problem of medical image segmentation and vendor-adaptation by unet-gan," in *MICCAI*, 2019, pp. 623–631.

[7] 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 *AAAI*, 2019, pp. 865–872.

[8] X. Liu, X. Wei, A. Yu, and Z. Pan, "Unpaired data based cross-domain synthesis and segmentation using attention neural network," in *Proceedings of The Eleventh Asian Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 101. Nagoya, Japan: PMLR, 17–19 Nov 2019, pp. 987–1000.

[9] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *CoRR*, vol. abs/1511.06434, 2015.

[10] C. Chen *et al.*, "Unsupervised multi-modal style transfer for cardiac mr segmentation," in *STACOM*, 2020, pp. 209–219.

[11] J. Hoffman *et al.*, "CyCADA: Cycle-consistent adversarial domain adaptation," in *ICML*, vol. 80, 10–15 Jul 2018, pp. 1989–1998.

[12] X. Zhuang *et al.*, "Cardiac segmentation on late gadolinium enhancement mri: A benchmark study from multi-sequence cardiac mr segmentation challenge," 2020.

[13] X. Zhuang *et al.*, "Evaluation of algorithms for multi-modality whole heart segmentation: An open-access grand challenge," *Medical Image Analysis*, vol. 58, p. 101537, 2019.

[14] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *ICCV*, pp. 2242–2251, 2017.

[15] P. Russo, F. M. Carlucci, T. Tommasi, and B. Caputo, "From source to target and back: Symmetric bi-directional adaptive gan," in *CVPR*, June 2018.

[16] Y. Zhang, S. Miao, T. Mansi, and R. Liao, "Task driven generative

modeling for unsupervised domain adaptation: Application to x-ray image segmentation," in *MICCAI*, 2018, pp. 599–607.

[17] H. Zhao, H. Li, S. Maurer-Stroh, Y. Guo, Q. Deng, and L. Cheng, "Supervised segmentation of un-annotated retinal fundus images by synthesis," *IEEE Transactions on Medical Imaging*, vol. 38, no. 1, pp. 46–56, 2019.

[18] P. Liu, B. Kong, Z. Li, S. Zhang, and R. Fang, "Cfea: Collaborative feature ensembling adaptation for domain adaptation in unsupervised optic disc and cup segmentation," in *MICCAI*, 2019, pp. 521–529.

[19] Y. Tsai, W. Hung, S. Schulter, K. Sohn, M. Yang, and M. Chandraker, "Learning to adapt structured output space for semantic segmentation," in *CVPR*, 2018, pp. 7472–7481.

[20] T. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, "Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *CVPR*, June 2019, pp. 2512–2521.

[21] 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 transactions on medical imaging*, vol. 38, no. 11, pp. 2485–2495, 2019.

[22] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *ECCV 2018*, 2018, pp. 179–196.

[23] Z. Shanis, S. Gerber, M. Gao, and A. Enquobahrie, "Intramodality domain adaptation using self ensembling and adversarial training," in *Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data*, 2019, pp. 28–36.

[24] Z. Zhang, L. Yang, and Y. Zheng, "Translating and segmenting multimodal medical volumes with cycle- and shape-consistency generative adversarial network," in *CVPR*, June 2018, pp. 9242–9251.

[25] P. Delisle, B. Anctil-Robitaille, C. Desrosiers, and H. Lombaert, "Adversarial normalization for multi domain image segmentation," in *ISBI*, 2020, pp. 849–853.

[26] K. Kamnitsas *et al.*, "Unsupervised domain adaptation in brain lesion segmentation with adversarial networks," in *Information Processing in Medical Imaging*, 2017, pp. 597–609.

[27] Q. Dou *et al.*, "Pnp-adanet: Plug-and-play adversarial domain adaptation network at unpaired cross-modality cardiac segmentation," *IEEE Access*, vol. 7, pp. 99 065–99 076, 2019.

[28] S. Wang, L. Yu, K. Li, X. Yang, C.-W. Fu, and P.-A. Heng, "Boundary and entropy-driven adversarial learning for fundus image segmentation," in *MICCAI*, 2019, pp. 102–110.

[29] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Perez, "Dada: Depth-aware domain adaptation in semantic segmentation," in *ICCV*, October 2019.

[30] 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 Transactions on Medical Imaging*, vol. 39, no. 7, pp. 2494–2505, 2020.

[31] J. Wang, H. Huang, C. Chen, W. Ma, Y. Huang, and X. Ding, "Multi-sequence cardiac mr segmentation with adversarial domain adaptation network," in *STACOM*, 2020, pp. 254–262.

[32] H. Roth, W. Zhu, D. Yang, Z. Xu, and D. Xu, "Cardiac segmentation of lge mri with noisy labels," in *STACOM*, 2020, pp. 228–236.

[33] J. Cai, Y. Xia, D. Yang, D. Xu, L. Yang, and H. Roth, "End-to-end adversarial shape learning for abdomen organ deep segmentation," in *International Workshop on Machine Learning in Medical Imaging*. Springer, 2019, pp. 124–132.

[34] S. Vesal, N. Ravikumar, and A. Maier, "Automated multi-sequence cardiac mri segmentation using supervised domain adaptation," in *STACOM*, 2020, pp. 300–308.

[35] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5967–5976.

[36] H. Fan, H. Su, and L. Guibas, "A point set generation network for 3d object reconstruction from a single image," in *CVPR*, 2017, pp. 2463–2471.

[37] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *CVPR*, 2017, pp. 77–85.

[38] S. Wang, L. Yu, K. Li, X. Yang, C.-W. Fu, and P.-A. Heng, "Boundary and entropy-driven adversarial learning for fundus image segmentation," in *MICCAI*. Springer, 2019, pp. 102–110.

[39] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015, pp. 234–241.

[40] J. Zhuang, Z. Chen, J. Zhang, D. Zhang, and Z. Cai, "Domain adaptation for retinal vessel segmentation using asymmetrical maximum classifier discrepancy," in *Proceedings of the ACM Turing Celebration Conference - China*, ser. ACM TURC '19, New York, NY, USA, 2019.

[41] B. Ly, H. Cochet, and M. Sermesant, "Style data augmentation for robust segmentation of multi-modality cardiac mri," in *STACOM*, 2020, pp. 197–208.

[42] X. Wang *et al.*, "Sk-unet: An improved u-net model with selective kernel for the segmentation of multi-sequence cardiac mr," in *STACOM*, 2020, pp. 246–253.

[43] V. M. Campello, C. Martín-Isla, C. Izquierdo, S. E. Petersen, M. A. G. Ballester, and K. Lekadir, "Combining multi-sequence and synthetic images for improved segmentation of late gadolinium enhancement cardiac mri," in *STACOM*, 2020, pp. 290–299.

[44] Y. Liu, W. Wang, K. Wang, C. Ye, and G. Luo, "An automatic cardiac segmentation framework based on multi-sequence mr image," in *STACOM*, 2020, pp. 220–227.

[45] J. Chen, H. Li, J. Zhang, and B. Menze, "Adversarial convolutional networks with weak domain-transfer for multi-sequence cardiac mr images segmentation," in *STACOM*, 2020, pp. 317–325.

[46] A. B. Jung *et al.*, "imgaug," https://github.com/aleju/imgaug, 2020, online; accessed 01-Feb-2020.

[47] R. Bermúdez-Chacón, P. Márquez-Neila, M. Salzmann, and P. Fua, "A domain-adaptive two-stream u-net for electron microscopy image segmentation," *ISBI*, pp. 400–404, 2018.

[48] O. Bernard *et al.*, "Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved?" *IEEE Transactions on Medical Imaging*, vol. 37, no. 11, pp. 2514–2525, 2018.