# Moment-Consistent Contrastive CycleGAN for Cross-Domain Pancreatic Image Segmentation

Zhongyu Chen, Yun Bian, Erwei Shen, Ligang Fan, Weifang Zhu, Fei Shi, Chengwei Shao, Xinjian Chen, *Senior Member, IEEE*, and Dehui Xiang, *Member, IEEE*

***Abstract*— CT and MR are currently the most common imaging techniques for pancreatic cancer diagnosis. Accurate segmentation of the pancreas in CT and MR images can provide significant help in the diagnosis and treatment of pancreatic cancer. Traditional supervised segmentation methods require a large number of labeled CT and MR training data, which is usually time-consuming and laborious. Meanwhile, due to domain shift, traditional segmentation networks are difficult to be deployed on different imaging modality datasets. Cross-domain segmentation can utilize labeled source domain data to assist unlabeled target domains in solving the above problems. In this paper, a cross-domain pancreas segmentation algorithm is proposed based on Moment-Consistent Contrastive Cycle Generative Adversarial Networks (MC-CCycleGAN). MC-CCycleGAN is a style transfer network, in which the encoder of its generator is used to extract features from real images and style transfer images, constrain feature extraction through a contrastive loss, and fully extract structural features of input images during style transfer while eliminate redundant style features. The multi-order central moments of the pancreas are proposed to describe its anatomy in high dimensions and a contrastive loss is also proposed to constrain the moment consistency, so as to maintain consistency of the pancreatic structure and shape before and after style transfer. Multi-teacher knowledge distillation framework is proposed to transfer the knowledge from multiple teachers to a single student, so as to improve the robustness and performance of the student network. The experimental results have demonstrated the superiority of our framework over state-of-the-art domain adaptation methods.**

***Index Terms*— Domain adaptation, moment-consistent, contrastive learning.**

## I. Introduction

PANCREATIC cancer has strong concealment, rapid onset, and high malignancy. Its five-year survival rate is only 5%-10%, and hence it is called the "king of cancers". According to the American Cancer Annual Report [1] in 2020, a total of 57,600 people were diagnosed with pancreatic cancer, and 47,050 people died of pancreatic cancer. Pancreatic cancer has become one of worldwide cancers, and early detection plays a crucial role in its treatment.

Computed Tomography (CT) and Magnetic Resonance (MR) are two common imaging techniques for pancreatic cancer [2]. The imaging principles of CT and MR result in large modality differences. CT images typically show uniform intensity distribution and strong contrast, with clear tissue boundaries and contours. In contrast, MR images usually have uneven intensity distribution and relatively indistinct tissue boundaries and contours. CT and MR have different imaging expressions for the pancreas, providing doctors with reliable data support for making accurate diagnosis from different modalities.

Deep neural networks have achieved great success in pancreas segmentation [3], [4], when the training and testing data are drawn from the same distribution. These works require a large amount of annotated data. However, acquiring such a large amount of data is time-consuming and labor-intensive, especially for medical images that require diagnostic expertise [5]. It has been pointed out that well-trained models often fail when tested on data from different modalities, as medical images acquired from different modalities have very different characteristics [6], [7], [8], [9]. Severe domain shift between modalities usually reduces the performance of well-trained deep neural networks. Although it is easy for human to

Manuscript received 21 July 2024; accepted 15 August 2024. Date of publication 21 August 2024; date of current version 2 January 2025. This work was supported in part by the National Nature Science Foundation of China under Grant 62371328, Grant 61971298, Grant 61771326, Grant 62271337, Grant U20A20170, Grant 81871352, Grant 82171915, Grant 82171930, Grant 82271972, and Grant 82371955; in part by the National Key Research and Development Program of China under Grant 2018YFA0701700; in part by the Natural Science Foundation of Shanghai Science and Technology Innovation Action Plan under Grant 21ZR1478500 and Grant 21Y11910300; and in part by the Clinical Research Plan of Shanghai Hospital Development Center (SHDC) under Grant SHDC2022CRD028. *(Zhongyu Chen and Yun Bian contributed equally to this work.) (Corresponding author: Dehui Xiang.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by Shanghai Changhai Hospital Ethics Committee under Application No. CHEC2021-163.

Zhongyu Chen, Erwei Shen, Ligang Fan, Weifang Zhu, Fei Shi, Xinjian Chen, and Dehui Xiang are with the School of Electronic and Information Engineering, Soochow University, Suzhou, Jiangsu 215006, China (e-mail: czy8785@gmail.com; yancheng15688@163.com; 1084640268@qq.com; wfzhu@suda.edu.cn; shifei@suda.edu.cn; xjchen@suda.edu.cn; xiangdehui@suda.edu.cn).

Yun Bian and Chengwei Shao are with the Department of Radiology, Changhai Hospital, Navy Military Medical University, Shanghai 200433, China (e-mail: bianyun2012@foxmail.com; cwshao@sina.com).

Digital Object Identifier 10.1109/TMI.2024.3447071

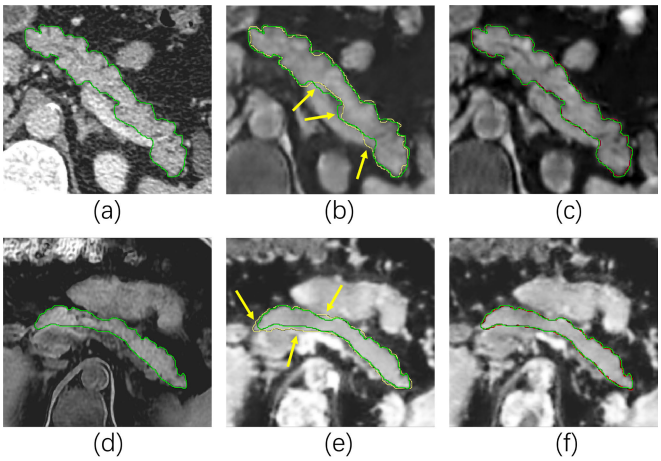


Fig. 1. Pancreas deformation in the image transformation process. Green lines represent ground truth of original images, yellow and red lines are manual annotations of the synthesized images generated by CycleGAN and our method, respectively. The first row shows CT to MR adaptation, and the second row shows MR to CT adaptation: a) Original CT image, b) synthesized MR image by CycleGAN, c) synthesized MR image by our method, d) Original MR image, e) synthesized CT image by CycleGAN, f) synthesized CT image by our method.

recognize the same anatomy across modalities, the deep neural networks trained on MR/CT data may fail in segmenting CT/MR images. Therefore, it is necessary to develop cross-modality image transformation methods to effectively transfer the knowledge learned from the source domain to the target domain without using additional annotations for the target domain.

Many researches on pattern recognition, image processing, and computer vision [10] are dedicated to transferring images from one domain to another. These methods require a set of paired images that are identical in terms of content and structure but differ in style. However, obtaining paired images is a challenging task, and sometimes even impractical to obtain. Recently, inspired by Generative Adversarial Network (GAN) [11], a large portion of studies [12], [13], [14] are proposed to achieve unsupervised domain adaptation in medical image analysis by image appearances and latent feature alignment without a paired data. For example, Cycle Adversarial Networks(CycleGAN) [12] is widely used to transform images cross modalities. However, CycleGAN cannot effectively preserve the structure and shape of the pancreas. As shown in Fig.1, the images transformed by CycleGAN exhibit significant differences in the shape and structure of the pancreas compared to the original images. Common domain adaptation methods are unable to achieve good results in our cross-domain pancreas segmentation task due to the following challenges: (1) Small size of the pancreas. The volume of the pancreas accounts for less than 1% of the entire CT and MR scan; (2) Low contrast. The edges of the pancreas are relatively blurred and the contrast is also low between the pancreas and its surrounding tissues and organs; (3) Significant anatomical differences in shape, size, structure, and position of the pancreas among different patients; (4) Significant style differences between CT and MR images.

To address above issues, a novel Moment-Consistent Contrastive Cycle Adversarial Networks(MC-CCycleGAN) is presented for unsupervised cross-domain pancreas segmentation. The proposed framework consists of two main subnetworks: (1) an image transformation subnetwork; (2) multi-teacher knowledge distillation subnetwork. The first subnetwork is used to transform the image from the source/target domain into the target/source domain. The second subnetwork takes the synthesized images as input to accomplish the segmentation tasks. In the first subnetwork, MC-CCycleGAN is proposed as the style transfer network, which introduces the contrastive learning within the framework of CycleGAN. The contrastive learning method is used to fully extract the structural features of the input image in the generator's encoder of MC-CCycleGAN, while eliminate redundant style features. This greatly improves MC-CCycleGAN's ability to perform image style transfer, reduces the style differences between different domain images, and ensures the consistency of image content. In order to maintain the consistency of the structure and shape of the pancreas in both the original and synthesized images, the multi-order central moments of the pancreas are computed to provide a high-dimensional description of its anatomy, and perform contrastive learning on the consistent moments. In the second subnetwork, a multi-teacher knowledge distillation framework is proposed for target domain pancreas segmentation. Multi-teacher models are pre-trained using synthesized images generated by MC-CCycleGAN from different training epochs. With this framework, the model is compressed by transferring the knowledge from multiple teachers to a single student model. Comprehensive experiments have been performed with one in-house dataset: Multi-Modality Pancreas Segmentation (MMPS) dataset, and two publicly available datasets: Abdominal Multi-Organ Benchmark for Versatile Medical Image Segmentation(AMOS) and Multi-Modality Whole Heart Segmentation (MMWHS) Challenge 2017. The experimental results have demonstrated the superiority of our framework over state-of-the-art domain adaptation methods. Our main contributions are summarized as,

1. CCycleGAN is proposed to preserve the structural features of images and remove excessive style features during image style transfer.

2. Multi-order central moments are integrated into CCycleGAN to describe its anatomy in high dimension and a contrastive loss is proposed to ensure the consistency of the multi-order central moments, so as to maintain consistency of the structure and shape before and after style transfer.

3. A multi-teacher knowledge distillation framework is proposed to transfer the knowledge from multiple teachers to a single student model.

## II. Related Work

Many pancreas segmentation methods [3], [4], [15], [16] have been proposed for CT and MR images. However, due to the huge difference between CT and MR images, these methods often failed in cross-domain (CT to MR or MR to CT) pancreas segmentation. Domain adaptation, which aimed to overcome the distribution difference of different domains, has attracted substantial research efforts, both in

computer vision [17], [18] and medical imaging [8], [19]. Many deep learning-based studies [14], [20], [21] have been dedicated to transferring the knowledge learned from the source domain to the target domain in a supervised or unsupervised way. However, obtaining labeled data was labor-intensive and time-consuming, unsupervised domain adaptation was more desirable. The current unsupervised domain adaptation methods could be divided into three classes: feature alignment based on difference measurement, domain feature disentanglement-based methods, and pixel-level and feature alignment based on GAN [11].

The feature alignment method based on difference measurement mitigated domain shift by shrinking the distribution differences from different domains in the feature space. Some early researches focused on reducing domain difference to achieve domain alignment. Maximum Mean Discrepancy (MMD) [22] was a common domain difference measurement used to assess the similarity between two distributions. Another common domain difference measurement was the distance between two distributions to obtain domain-invariant features for the final segmentation task [23].

Feature disentanglement was employed to solve unsupervised domain adaptation [24], [25], [26]. The goal of learning disentangled representations was to develop a model that can represent the distinct factors in the data. Chang et al. [25] tried to disentangle images into domain-invariant structure and domain-specific texture representations. Xie et al. [26] first factorized an image into domain-invariant anatomy and domain-specific modality components by disentanglement learning, and then utilized self-training strategy to further improve the segmentation performance.

Adversarial learning was also used to mitigate domain shift from different perspectives, including image-level alignment, feature-level alignment, and their combinations. The advent of GAN [11] proposed image alignment methods that transformed source/target images into target/source-like images. Bi et al. [27] aimed to synthesize positron emission tomography images from CT images via multi-channel generative adversarial networks. Dar et al. [28] utilized conditional GAN for multi-contrast MR synthesis. The success of CycleGAN inspired many image alignment methods to further regularize the image transformation process with additional constrains. Zhu et al. [12] proposed CycleGAN to perform unpaired image-to-image transformation with cycle consistency loss to preserve the structure. Jiang et al. [29] utilized CycleGAN to synthesize MR images from CT images, and then the synthesized MR images were combined with a few real MR data for semi-supervised tumor segmentation. Huo et al. [30] proposed an end-to-end synthesis and segmentation network (SSNet) that CycleGAN and a segmentation network were integrated to segment synthesized images. Though these methods achieved good results and could generate high-quality synthesized images, they did not impose semantic constraints, which could not guarantee the consistency of structure during the transformation. Tomar et al. [31] imposed the auxiliary semantics to handle the geometric changes and preserve structures during image transformation. Meanwhile, other studies focused on feature alignment to extract domain-invariant features by adversarial learning. Ganin et al. [32] tried to perform adversarial learning in the feature space to differentiate the features across domains. To project the high-dimensional feature space to other compact spaces, Tsai et al. [18] extended adversarial learning to the semantic prediction space. Although alignments on image or feature level achieved good results in unsupervised domain adaptation, the combination of these two techniques achieved a stronger domain adaption performance. Hoffman et al. [33] enforced cycle-consistency and a task loss. Zhang et al. [34] explored domain adaptation for semantic segmentation from the viewpoint of both visual appearance-level and representation-level adaptation. However, their image and feature alignments were connected in a sequential manner and trained separately at different stages without any interaction. To fully explore the simultaneous alignments from the feature alignment and image alignment, Chen et al. [21] performed bidirectional unsupervised domain adaptation between cardiac CT and MR images by conducting synergistic alignment of domains.

## III. Method

An unsupervised domain adaptation framework is proposed for pancreas segmentation based on MC-CCycleGAN. The framework consists of image transformation subnetwork, and multi-teacher knowledge distillation subnetwork. As shown in Fig.2, the image transformation subnetwork is designed to translate images between the source domain and the target domain, and then the synthesized images are used for the subsequent segmentation subnetwork. The details are introduced in the following sections.

### A. Image Transformation Subnetwork

Different modality data exhibit various visual characteristics as a result of differences in their distribution. Cross-modality data alignment can be performed by mutual transfer between modalities. The object's spatial and structural information is maintained in the process of the modality transfer. Therefore, domain invariant features can be extracted between different modalities to address the issue of unlabeled target domain segmentation. Formally, given an annotated source domain datataset $X^S : \{x_i^s, y_i^s\}_{i=1}^{N_s}$ ($x_i^s$ is the $i$th annotated source domain image, $y_i^s$ is the label of $x_i^s$, $N^s$ is the number of the annotated source domain images) and an unannotated target domain dataset $X^T : \{x_i^t\}_{i=1}^{N_t}$ ($x_i^t$ is the $i$th unannotated target domain image, $N^t$ is the number of the unannotated target domain images). The source domain data $X^S$ are transferred to the target domain data $X^{S\to T}$. The synthesized target domain images should appear the similar style of the target domain, while preserve the original contents and structural semantics without any change of the source domain.

Inspired by CycleGAN [12], a generator $G_t$ and a discriminator $D_t$ are constructed to transfer images between source and target domain. The generator is used to transfer the source images to target-like images, namely $x^{s\to t} = G_t(x^s)$. The discriminator is trained to correctly distinguish between the real target image $x^t$ and synthesized target image $x^{s\to t}$. $G_t$ and $D_t$ are optimized by the adversarial loss $L_{adv}^t(G_t, D_t)$.

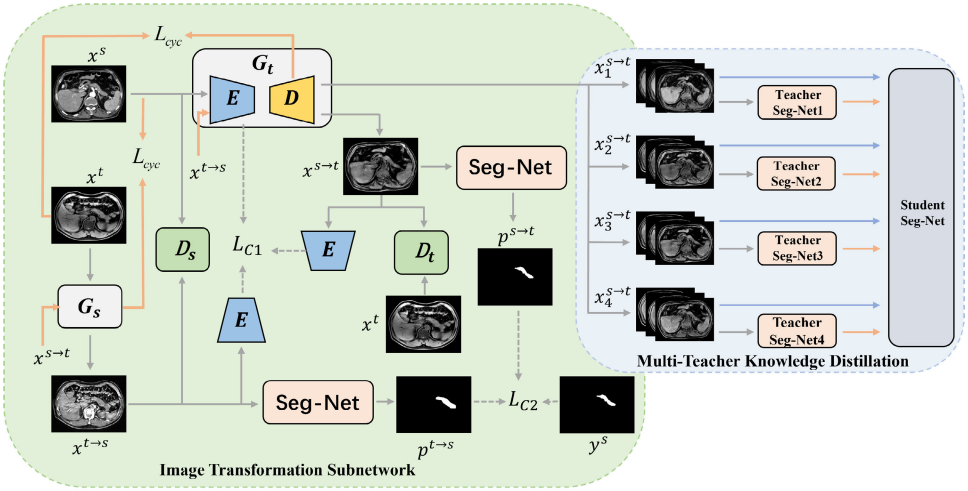


Fig. 2. The framework of our method for unsupervised domain adaptation. The generator $G_t$ and $G_s$ serve the source-to-target and target-to-source image transformation, while $D_t$ and $D_s$ are their corresponding discriminators. The encoder $E$ is used to encode the image for contrastive learning. Seg-Net is used to produce the segmentation prediction of the synthesized images, and the predictions are used for the calculation of multi-order central moments. $x_1^{s\to t}$ to $x_4^{s\to t}$ represent the synthesized images selected from different training epochs to train the Multi-Teacher Knowledge Distillation network. These synthesized images have similar pancreatic structures but minor stylistic differences. $p^{s\to t}$ and $p^{t\to s}$ are the prediction maps of synthesized images $x^{s\to t}$ and $x^{t\to s}$. $y^s$ is the label of a source domain image.

To synthesize source-like images, a reverse generator $G_s$ and discriminator $D_s$ are also constructed. The generation of a synthesized source image is $x^{t\to s} = G_s(x^t)$. $G_s$ and $D_s$ are optimized by the adversarial loss $L_{adv}^s(G_s, D_s)$. In order to maintain the content of the synthesized image consistent with the original image throughout the transfer process, a reverse generator is used on the synthesized images. The reconstruction of $x^{s\to t}$ back to the source domain is $x^{s\to t\to s} = G_s(G_t(x^s))$. The target domain reconstruction is similar to the source domain reconstruction, $x^{t\to s\to t} = G_t(G_s(x^t))$. The overall reconstruction loss is defined as $L_{cyc}(G_t, G_s)$. In addition, Seg-Net is also constructed to generate the predictions of synthesized images which will be used to calculate the multi-order central moments. The trained Seg-Net from different epochs will be regarded as the pretrained teacher networks.

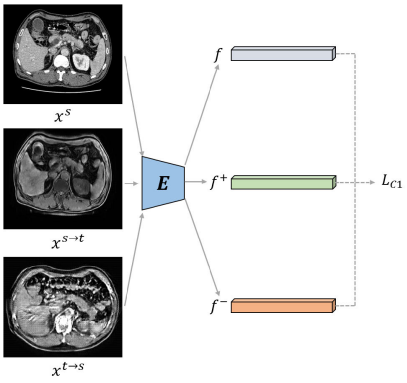


Fig. 3. The contrastive loss $l_{c1}$ for feature alignment. The encoder $E$ in generator $G_t$ is used to encode,the query, positive and negative samples and generate corresponding feature maps. Afterwards, these feature maps pass through fully connected layers to obtain the desired latent variables, which are then used to compute the contrastive loss.

## B. Contrastive Learning for Feature Alignment

CycleGAN has been able to generate fake images by introducing adversarial learning at the image level. However, when the domain shift is large, the encoder of the generator in CycleGAN cannot fully extract the structural and style features from the original image, which tends to result in the inconsistency of the structure between the synthesized image and the original image. To alleviate the problem, the encoder of the generator is aimed to only extract the structural features of the image while discard redundant style features, in order to maintain consistency of the structure and content between the synthesized image and the original image.

As shown in Fig.2, $x^{s\to t}$ is the target-like image transferred from $x^s$, and its content and structure should be consistent with $x^s$. $x^{t\to s}$ is the source-like image transferred from $x^t$, and it has a similar style and appearance with $x^s$, but the content and structure of $x^{t\to s}$ are largely different from those of $x^s$. The selection of negative samples in our method is revised for traditional contrastive learning methods [35]. In previous methods, only the remaining data in the dataset after selecting good positive samples were used as negative samples. This can narrow the gap between positive and negative samples since there is still similarity between them and it tends to greatly reduce the effectiveness of contrastive learning. In our work, $x^s$ and $x^{s\to t}$ are respectively regarded as query and its "positive" sample, and $x^{t\to s}$ is regarded as "negative" sample. The encoder $E$ in the generator $G_t$ is used as feature extractor. The query, positive, and negative samples are encoded by $E$ to obtain feature maps $fea^s$, $fea^{s\to t}$ and $fea^{t\to s}$. The feature maps are fed into the fully connected layer $f_c$ to obtain hidden variables $z^s$, $z^{s\to t}$ and $z^{t\to s}$. InfoNCE loss is employed to reduce the distance between positive sample pairs and increase the distance between negative sample pairs. As shown in Fig.3, a contrastive loss can be defined as,

$$l_{c1}\left(z^s, z^{s\to t}, z^{t\to s}\right) = \log\left(1+\exp\left(\left(sim\left(z^s, z^{t\to s}\right) - sim\left(z^s, z^{s\to t}\right)\right)/\tau\right)\right), \quad (1)$$

where $\tau$ represents the temperature coefficient, which is used to scale the distance between the query and other samples. The sim function is used to calculate the similarity between vectors, where a larger value indicates greater similarity. For two $N^z$-dimensional vectors $u \in \mathbb{R}^{N^z}$ and $v \in \mathbb{R}^{N^z}$, their cosine similarity can be used to measure the similarity between the two vectors as

$$cos(\theta) = \frac{\sum_{i=1}^{N^z} (u_i \times v_i)}{\sqrt{\sum_{i=1}^{N^z} (u_i)^2} \times \sqrt{\sum_{i=1}^{N^z} (v_i)^2}} \tag{2}$$

The contrastive loss $l_{c1}$ is used to encourage $E$ to capture domain-invariant features such as structure and content, and discard style features and other domain-specific features.

### C. Moment Consistency

The pancreas has strong anatomical specificity, with a small and irregular shape, and fuzzy edges. This makes it difficult to maintain structural consistency in the process of image transformation. The image central moments are translational invariant and can be used to describe shape features. Central moments are a set of mathematical indicators used to describe the shape of an object. The multi-order central moments [36] provide a high-dimensional description of the structure and shape of an object. Therefore, multi-order moments are introduced to help the constraint of structure. The multi-order central moments are defined as,

$$\begin{aligned}
\pi_1 &= M_{11} - \overline{x} M_{01} = M_{11} - \overline{y} M_{10}, \\
\pi_2 &= M_{20} - \overline{x} M_{10}, \\
\pi_3 &= M_{02} - \overline{y} M_{01}, \\
\pi_4 &= M_{21} - 2\overline{x} M_{11} - \overline{y} M_{20} + 2\overline{x}^2 M_{01}, \\
\pi_5 &= M_{12} - 2\overline{y} M_{11} - \overline{x} M_{02} + 2\overline{y}^2 M_{10}, \\
\pi_6 &= M_{30} - 3\overline{x} M_{20} + 2\overline{x}^2 M_{10}, \\
\pi_7 &= M_{03} - 3\overline{y} M_{02} + 2\overline{y}^2 M_{01},
\end{aligned} \tag{3}$$

where $(\overline{x}, \overline{y})$ denotes the centroid of the image, $M_{ij}$ is the raw moment of the image $I(x, y)$ as,

$$\begin{aligned}
M_{ij} &= \sum_x \sum_y x^i y^j I(x, y), \\
\overline{x} &= \frac{M_{10}}{M_{00}}, \overline{y} = \frac{M_{01}}{M_{00}}.
\end{aligned} \tag{4}$$

The corresponding multi-order central moments are different if shape deformation is produced in the process of pancreas segmentation. Therefore, central moments and the contrastive learning are combined to constrain the anatomical consistency in the image transformation.

As shown in Fig.2, a segmentation subnetwork is built to segment $x^{s\to t}$ and $x^{t\to s}$, and get the prediction $p^{s\to t}$ and $p^{t\to s}$. $p^{s\to t}$ should have the same structure as $y^s$, while $p^{t\to s}$ is different, and thus $p^{s\to t}$ and $y^s$ are respectively regarded as query and its "positive" sample, and $p^{t\to s}$ is regarded as "negative" sample. These positive and negative sample pairs reflect the shape and structure of the pancreas. When using a fully connected layer to reduce dimensionality, the generated latent variables may lose much shape and structural information of the pancreas, which may lead to uselessness of the contrastive loss. To solve this problem, a multi-order central moment vector $\Pi = [\pi_1, \pi_2, \pi_3, \pi_4, \pi_5, \pi_6, \pi_7]$ is respectively used for $p^{s\to t}$, $y^s$ and $p^{t\to s}$ and the calculated multi-order central moment vectors are respectively considered as latent variables $a^{s\to t}$, $a^s$ and $a^{t\to s}$ to abstractly represent the shape and structure of the pancreas. The higher the similarity between the latent variables, the more similar the corresponding predicted results are, and therefore, as shown in Fig.4, a central moment consistency loss can be defined as,

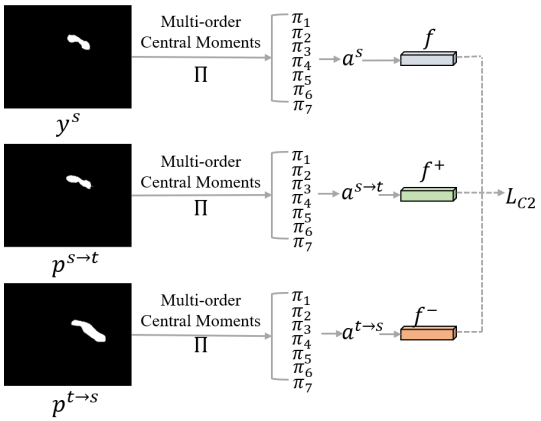


Fig. 4. The moment consistency loss $l_{c2}$.

$$\begin{aligned}
&l_{c2}\left(a^s, a^{s\to t}, a^{t\to s}\right) = \\
&\log\left(1 + \exp\left(\left(sim\left(a^s, a^{t\to s}\right) - sim\left(a^s, a^{s\to t}\right)\right)/\tau\right)\right),
\end{aligned} \tag{5}$$

where cosine similarity is also used as the $sim$ function to calculate the similarity between two vectors. By optimizing the loss to minimize the distance between positive sample pairs and maximize the distance between negative sample pairs, consistency in pancreatic structure and shape are maintained before and after style transfer.

### D. Multi-Teacher Knowledge Distillation

As training epoch of image transformation subnetwork increases, the synthesized image $x^{s\to t}$ is progressively close to the target image $x^t$. The multiple styles of $x^{s\to t}$ enrich the diversity of training samples for target domain image segmentation. Therefore, a multi-teacher knowledge distillation subnetwork is proposed to improve the target domain image segmentation. The synthesized images from different training epochs are used to train $N_e$ teacher networks, respectively. In the implementation, $N_e$ is set to 4, as shown in Fig.2. The teacher networks are then used to guide the training of the student network. The predicted vector produced by the $j$th teacher network for an input image $x_j^{s\to t}$ is represented by $(z_r)_j^{s\to t} \in \{(z_r)_1^{s\to t}, \ldots, (z_r)_{N_e}^{s\to t}\}$. By introducing the temperature $\tau$, the sigmoid layer converts the predicted vector $(z_r)_j^{s\to t}$ into a probability distribution $(p_r^\tau)_j^{s\to t} \in \{(p_r^\tau)_1^{s\to t}, \ldots, (p_r^\tau)_{N_e}^{s\to t}\}$ as,

$$(p_r^\tau)_j^{s\to t} = sigmoid((z_r)_j^{s\to t}/\tau) \tag{6}$$

The synthesized images from the $N_e$ epochs and corresponding source domain labels will also be used for the training

of student network. The predicted vector produced by the student network for an input image $x_j^{s\to t}$ is represented by $(z_u)_j^{s\to t} \in \{(z_u)_1^{s\to t}, \ldots, (z_u)_{N_e}^{s\to t}\}$. The sigmoid layer converts the logits vector $(z_u)_j^{s\to t}$ to a probability distribution $(p_u^\tau)_j^{s\to t} \in \{(p_u^\tau)_1^{s\to t}, \ldots, (p_u^\tau)_{N_e}^{s\to t}\}$. $(p_u^\tau)_j^{s\to t}$ is computed by sigmoid with the same temperature $\tau$ as the teacher model,

$$(p_u^\tau)_j^{s\to t} = sigmoid((z_u)_j^{s\to t}/\tau) \tag{7}$$

Two loss functions are defined to train the student network. The first loss function $L_{s1}$ consists of binary cross-entropy loss and Dice loss,

$$L_{s1} = -\frac{1}{M}\left(\sum_{k=1}^{M} y_k^s \log (p_u^\tau)_{jk}^{s\to t} + \frac{\sum_{k=1}^{M} 2(p_u^\tau)_{jk}^{s\to t} y_k^s}{\sum_{k=1}^{M} ((p_u^\tau)_{jk}^{s\to t})^2 + (y_k^s)^2}\right) \tag{8}$$

where $(p_u^\tau)_{jk}^{s\to t}$ denotes the $k$th pixel in $(p_u^\tau)_j^{s\to t}$ and $y_k^s$ denotes the $k$th pixel in ground truth $y^s$ of the source image $x^s$. $M$ is the number of pixels in $y^s$.

The probabilities produced by the teacher networks are used as "soft labels" for training the student network. The second loss function $L_{s2}$ is defined by the soft labels $(p_r^\tau)_j^{s\to t}$ and the prediction $(p_u^\tau)_j^{s\to t}$ produced by the student network.

$$L_{s2} = -\frac{1}{M}\sum_{k=1}^{M} (p_r^\tau)_{jk}^{s\to t} \log (p_u^\tau)_{jk}^{s\to t}, \tag{9}$$

where $(p_r^\tau)_{jk}^{s\to t}$ denotes the $k$th pixel in $(p_r^\tau)_j^{s\to t}$.

The total loss $L_{total}$ is defined as,

$$\begin{aligned} L_{total} = {} & \lambda_{adv}^t L_{adv}^t (G_t, D_t) + \lambda_{adv}^s L_{adv}^s (G_s, D_s) \\ & + \lambda_{cyc} L_{cyc} (G_t, G_s) + \lambda_{c1} l_{c1} \left(z^s, z^{s\to t}, z^{t\to s}\right) \\ & + \lambda_{c2} l_{c2} \left(a^s, a^{s\to t}, a^{t\to s}\right) + \lambda_{s1} L_{s1} + \lambda_{s2} L_{s2} \end{aligned} \tag{10}$$

where $\lambda_{adv}^t, \lambda_{adv}^s, \lambda_{cyc}, \lambda_{c1}, \lambda_{c2}, \lambda_{s1}, \lambda_{s2}$ are the balance parameters of corresponding losses.

## IV. Experiments

### A. Datasets

Our proposed unsupervised domain adaptation method is evaluated on three datasets: Multi-Modality Pancreas Segmentation (MMPS) private dataset, Abdominal Multi-Organ Benchmark for Versatile Medical Image Segmentation(AMOS) and Multi-Modality Whole Heart Segmentation (MMWHS) Challenge 2017.

*1) MMPS Dataset:* This dataset was collected from Changhai Hospital of Navy Medical University, and included 3D abdominal CT volume images of 97 pancreatic cancer patients and 3D abdominal MR volume images of 97 pancreatic cancer patients. Note that the CT and MR images were from different patients and they were unpaired. The images in MMPS dataset was based on transverse sections of both CT and MR images, with a slice size of 512 × 512 for CT images and two different sizes (512 × 512 and 320 × 260) for MR images. To maintain consistency in image size within the dataset, all images were uniformly downsampled to a size of 256 × 256. Due to the differences in the acquisition methods of CT and MR images, as well as individual difference between patients, the numbers of 2D axial CT images for different patients ranged from 30 to 100, while the numbers of 2D axial MR images for different patients ranged from 20 to 50. The ranges of spacings on the sagittal, coronal, and axial plane were [0.34,0.84]mm, [0.34,0.84]mm, and [0.25,1.0]mm for CT images and [0.63,1.73]mm, [0.63,1.73]mm, [2.0,6.0]mm for MR images, respectively. All data were accurately manually annotated under the guidance of professional doctors. The CT/MR images were respectively divided into three subsets with independent patients for three-fold cross validation. Each subset of CT images consisted of 32, 32 and 33 patients, respectively, while each subset of MR images contained 32, 32 and 33 patients, respectively. The data from the i-th (i=1,2,3) fold was used as the evaluation set, and the data from the remaining two folds were used as the training set. There was no subject overlap among the subsets. Three models were trained. The evaluation metrics of the three models on their respective evaluation sets were averaged as the final results. In addition, data augmentation techniques were applied, such as random horizontal and vertical flipping and random rotation within the range of [-10, 10] degrees to both CT and MR images.

*2) AMOS Dataset:* AMOS dataset was a public abdominal multi-organ dataset which contained the pancreas. It provided 500 CT and 100 MR scans collected from multi-center, multi-vendor, multi-modality, multi-phase, multi-disease patients. Only the training set and validate set were available, i.e. labels of only 300 CT and 60 MR scans were opened. To evaluate the pancreas segmentation performance of our proposed method, we focused on the pancreas segmentation. CT images were available in two sizes: 768 × 768 and 512 × 512, whereas the sizes of MR images varied between 60 × 320 and 468 × 576. The numbers of 2D axial CT images for different patients ranged from 68 to 353, while the numbers of 2D axial MR images for different patients ranged from 64 to 512. The ranges of spacings on the sagittal, coronal, and axial plane were [0.45,1.07]mm, [0.45,1.07]mm, and [1.25,5.0]mm for CT images and [0.69,1.95]mm, [0.69,3.0]mm, [0.82,3.0]mm for MR images, respectively. Detailed description about AMOS dataset can be found in [37]. All images were resized to 256 × 256. Similarly, the CT/MR images in AMOS dataset were also respectively divided into three subsets with independent patients for three-fold cross validation. Each subset of CT images consisted of 101, 101 and 98 scans, respectively, while each subset of MR images contained 22, 19 and 19 scans, respectively. The data from the i-th (i=1,2,3) fold was used as the evaluation set, and the data from the remaining two folds were used as the training set. There was no subject overlap among the subsets. Three models were trained. The evaluation metrics of the three models on their respective evaluation sets were averaged as the final results. Data augmentation techniques such as random horizontal and vertical flipping and random rotation within the range of [−10, 10] degrees were also applied to both CT and MR images.

*3) MMWHS Dataset:* MMWHS dataset consisted of 20 MRI and 20 CT whole cardiac images with ground truth. Note that the images across modalities were collected from different patients and they were unpaired. In our experiments, the ascending aorta (AA), the left atrium blood cavity (LAC), the left ventricle blood cavity (LVC), and the myocardium of the left ventricle (MYO) were segmented. To maintain consistency in image size within the dataset, all images were also resized to 256 × 256. Each sample image was normalized to have zero mean and unit variance in terms of the intensity value. For a fair comparison, the division of MR/CT images followed the previous works [21], [31], [38], [39], [40], and we randomly split each modality of the data into 80% training (16 subjects) and 20% testing (4 subjects) subsets for all experiments, so there was no subject overlap among the subsets. One model was trained. The evaluation metrics of the 4 subjects were averaged as the final results. Data augmentation techniques such as random horizontal and vertical flipping and random rotation within the range of $[-10, 10]$ degrees were also applied to both CT and MR images.

## B. Implementation Details

MC-CCycleGAN consisted of $G_s$, $G_t$, $D_s$ and $D_t$. The generators $G_s$ and $G_t$ had the same structure, and they both consisted of 3 convolutional layers, 3 deconvolutional layers, and followed by one convolutional layer to get the generated images. The discriminators $D_s$ and $D_t$ consisted of 5 convolutional layers and performed global average pooling to produce a single value output between 0 and 1. Teacher models and the student model were U-Net [41].

All networks were trained using Pytorch deep learning framework and accelerated by an NVIDIA GeForce RTX 3090 graphics card with a memory size of 24G. $\tau$ was set to 0.07. The balancing parameters $\lambda^t_{adv}, \lambda^s_{adv}, \lambda_{cyc}, \lambda_{c1}, \lambda_{c2}, \lambda_{s1}, \lambda_{s2}$ were set to 1, 1, 10, 1, 1, 1, 1, respectively. The whole network was trained separately. In stage 1, for the image transformation subnetwork, the Adam optimization algorithm was used with a learning rate (lr=0.0001), a first moment estimation exponential decay rate (beta1 = 0.5), a second moment estimation exponential decay rate (beta2 = 0.999). There were a total of 60 epochs deployed for its training and the batch size was set to 1 when using instance normalization [12]. In stage 2, for the multi-teacher knowledge distillation networks, the SGD optimization algorithm was used with a momentum optimizer set to 0.9 and a weight decay regularization coefficient was set to 0.0001. The learning rate was set to 0.01. The synthesized images generated in the 45th, 50th, 55th, and 60th training epochs of the image transformation subnetwork training were used to train the student network. The parameters of Seg-Net from the selected epochs were used as pretrained weights for the teacher networks. The total training epochs were 100, and the batch size was set to 4 when using batch normalization.

## C. Evaluation Metrics

Dice similarity coefficient (DSC), Jaccard similarity coefficient(Jac), Precision, Recall, average symmetric surface distance(ASD) and 95% Hausdorff Distance ($HD_{95}$) were used to quantitatively measure the performance of domain adaptation models for the segmentation task. Two-tailed Wilcoxon test of DSC was performed to compare the difference between our method and related methods, and $p$ <0.05 was considered to be statistically significant.

TABLE I
ABLATION EXPERIMENTS FOR CT->MR AND MT->CT DOMAIN ADAPTATION FOR MMPS

| Method | Pancreas CT→MR(%) | | | | Pancreas MR→CT(%) | | | |
|---|---|---|---|---|---|---|---|---|
| | Recall | Precision | Jac | DSC | Recall | Precision | Jac | DSC |
| CycleGAN | 73.24 | 69.15 | 53.77 | 66.63 | 66.03 | 71.28 | 51.33 | 64.11 |
| CCycleGAN | 77.81 | 75.26 | 60.75 | 73.09 | **74.04** | 68.32 | 54.91 | 67.38 |
| MC-CCycleGAN | 79.29 | 76.84 | 63.45 | 74.92 | 73.87 | 71.38 | 56.47 | 68.91 |
| MC-CCycleGAN+KD | **80.61** | **77.66** | **64.82** | **76.52** | 73.02 | **74.53** | **57.87** | **70.04** |

## D. Ablation Studies

Ablation studies were performed to evaluate the CCycleGAN, multi-order central moment consistency, and multi-teacher knowledge distillation. Table I presented the ablation studies for MMPS dataset of our key components. CycleGAN was regarded as the baseline to better measure the improvements of different components.

*1) CCycleGAN:* By imposing the constraint of feature contrastive learning through CCycleGAN, the positive samples are consistent with the query samples in structure and shape but different from styles while the negative samples are not consistent in structure and shape with the query samples although they have the similar style. Therefore, the encoder tends to gradually remove the style at the feature level, capture domain-invariant features and keep the consistency of structure and shape in the training of style transfer. As shown in second row of Table I, for the CT to MR adaptation, CCycleGAN respectively improved the Recall, Precision, Jac and DSC scores by 4.57, 6.11, 6.98 and 6.46 over CycleGAN, respectively. The introduction of contrastive learning for feature alignment enhances the contrast and structural features of the pancreas through the alignment of the structure and shape in synthetized target domain images and source domain images in the feature level. The improved Recall and Precision scores showed that more pancreatic tissues were segmented and less non-pancreatic tissues were incorrectly over-segmented. For the MR to CT adaptation, CCycleGAN improved the Recall, Jac and DSC scores by 8.01, 3.58 and 3.27 over CycleGAN, respectively. The highly improved Recall score showed that synthetized CT images were highly enhanced and more pancreatic tissues were segmented, although the Precision score was slightly decreased. Kindly note that the Jac and DSC scores were both improved, which showed that feature contrastive learning was beneficial to mitigate domain shift.

*2) Moment Consistency:* Multi-order central moments are used to describe the structure and shape of the pancreas. A contrastive loss is also deployed to decrease the differences of multi-order central moments, i.e. to maintain the consistency of pancreatic structure and shape. The multi-order central moments are integrated into contrastive learning

to align the pancreatic structure and shape in the image level. Based on CCycleGAN, the performance improvement of MC-CCycleGAN is also shown in third row of Table I. For the CT to MR adaptation, MC-CCycleGAN improved the Recall, Precision, Jac and DSC scores by 1.48, 1.58, 2.70 and 1.83 over CCycleGAN, respectively. These results demonstrated the strength of moment consistency in the proposed framework. The moment contrastive loss aligned the structure and shape of the pancreas in the synthetized MR images to ground truth of the CT images and differentiate the structure and shape of the pancreas in the synthetized MR images from the segmented pancreas in the synthetized CT image. More pancreatic tissues were segmented and less non-pancreatic tissues was over-segmented. For the MR to CT adaptation, MC-CCycleGAN improved the Precision, Jac and DSC scores by 3.06, 1.56 and 1.53 over CCycleGAN, respectively. The improved Precision score showed that synthetized CT images were highly improved and less non-pancreatic tissues was over-segmented, although the Recall score was slightly decreased.

*3) Multi-Teacher Knowledge Distillation:* The synthetized images can greatly enrich the training samples for student network training, since the structure and content of the images are very similar, and only minor differences exist in style and texture in the last training epochs, and therefore, the four teacher networks are built to jointly guide the training of a single student network, resulting in better segmented results. In Table I, the performance of our method has been further improved. For the CT to MR adaptation, knowledge distillation (KD) improved the Recall, Precision, Jac and DSC scores by 1.32, 0.82, 1.37 and 1.60 over MC-CCycleGAN, respectively. The improved Precision and Recall scores showed that KD could improve under-segmentation and restrain over-segmentation. For the MR to CT adaptation, KD improved the Precision, Jac and DSC scores by 3.15, 1.40 and 1.13 over MC-CCycleGAN, respectively. The improved Precision score showed that less non-pancreatic tissues was over-segmented, although the Recall score was slightly decreased. Kindly note that the Jac and DSC scores were both improved, which showed that KD was beneficial to improve target domain image segmentation.

The number of teacher networks was also worth exploring, as it directly affects whether the student network can receive effective guidance. Another ablation study had been deployed to research how the number of teacher networks influences the peformance of student network. As shown in Table VI, when using 1, 2, 4, and 6 teacher networks, the performance of student network firstly improved and then tended to stabilize. Therefore, using four teacher networks was reasonable.

*4) Hyperparameter Tuning:* To further investigate the impact of hyperparameters on network performance, several experiments were conducted with different $\lambda_{c1}$ and $\lambda_{c2}$. When $\lambda_{c1}$ and $\lambda_{c2}$ were individually set to 0.1, 0.5, 5 and 10, the results were shown in Table II. It indicated that inappropriate parameters can degrade network performance, and it was reasonable to set $\lambda_{c1}$ and $\lambda_{c2}$ to 1.

TABLE II
HYPERPARAMETERS $\lambda_{c1}$ AND $\lambda_{c2}$ TUNING

| value | $\lambda_{c1}$ | | | | $\lambda_{c2}$ | | | |
|---|---|---|---|---|---|---|---|---|
| | CT→MR | | MR→CT | | CT→MR | | MR→CT | |
| | ASD(mm) | DSC(%) | ASD(mm) | DSC(%) | ASD(mm) | DSC(%) | ASD(mm) | DSC(%) |
| 0.1 | 3.3 | 74.46 | 5.1 | 68.18 | 3.0 | 74.61 | 5.2 | 67.86 |
| 0.5 | 4.6 | 73.08 | 4.8 | 68.76 | 2.4 | 75.36 | 4.7 | 69.51 |
| 1 | **1.4** | **76.52** | **2.3** | **70.04** | **1.4** | **76.52** | **2.3** | **70.04** |
| 5 | 3.4 | 75.28 | 4.8 | 69.03 | 3.4 | 74.68 | 2.9 | 68.45 |
| 10 | 2.4 | 75.55 | 5.3 | 67.75 | 2.5 | 73.90 | 3.7 | 69.94 |

## *E. Comparisons With State-of-the-Art Methods in MMPS Dataset*

To observe the degradation of the segmentation caused by domain shift, segmentation performance of supervised training for target domain were respectively tested. "w/o adaptation" denoted that the model trained in source/target domain were directly applied to target/source images without using any domain adaptation.

Table III showed the segmented results of unsupervised domain adaptation (CT→MR) and (MR→CT) methods for the MMPS dataset. The supervised model for MR images obtained 83.44, 82.28, 70.33 and 80.79 in terms of the Recall, Precision, Jac and DSC scores, respectively. ASD was 1.2. The model trained on CT images and directly tested on MR images obtained 23.29, 79.01, 20.94 and 27.41 in terms of the Recall, Precision, Jac and DSC scores, respectively. ASD increased to 10.1. The supervised model for CT images obtained 85.11, 83.55, 72.78 and 82.63 in terms of the Recall, Precision, Jac and DSC scores, respectively. ASD was 1.1. The model trained on MR images and directly tested on CT images obtained 54.8, 50.86, 35.65 and 46.62 in terms of the Recall, Precision, Jac and DSC scores, respectively. ASD also increased to 6.5.

The significant performance gap between the "w/o adaptation" and the supervised training shows that there exists the severe domain shift between CT and MR images, which resulted in corresponding performance degradation of deep neural network on cross-modality segmentation tasks. Our proposed MC-CCycleGAN outperformed state-of-the-art domain adaptation methods, and achieved great improvements across different modalities. For the MR images, the average Recall, Jac and DSC score improved to 80.61, 64.82 and 76.52 over the pancreas segmentation, which were very close to the supervised training. For the CT images, the average Recall, Jac and DSC score improved to 73.02, 57.87 and 70.04. The qualitative results validated the effectiveness of our method on mitigating the severe domain shift. $p < 0.001$ of the DSC score shows that the superiority of our method for pancreas segmentation was statistically significant.

To validate the effectiveness of our method, different state-of-the-art unsupervised domain adaptation approaches were compared with the MMPS dataset. These unsupervised domain adaptation methods included RAM-DSIR [44], PLACE [43], VAE [42], AdaptPatch [47], AdaptSeg [18], PnP-AdaNet [45], Advent [46], Synseg [13], CycleGAN [12], CyCADA [33], SASAN [31], Prior SIFA [38], SIFA [21], DDFseg [48], UESM [40], DSAN [39], and Segment Anything Model (SAM) [49]. The visual comparison of these segmented results

TABLE III
PERFORMANCE COMPARISON WITH DIFFERENT UNSUPERVISED DOMAIN ADAPTATION METHODS FOR MMPS

| Method | CT→MR | | | | | | MR→CT | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Recall (%) | Precision (%) | $HD_{95}$ (mm) | ASD (mm) | Jac (%) | DSC (%) | Recall (%) | Precision (%) | $HD_{95}$ (mm) | ASD (mm) | Jac (%) | DSC (%)↑ |
| Supervised training | 83.44 | 82.28 | 9.9 | 1.2 | 70.33 | 80.79 | 85.11 | 83.55 | 10.2 | 1.1 | 72.78 | 82.63 |
| w/o adaptation | 23.29 | 79.01 | 26.0 | 10.1 | 20.94 | 27.41 | 54.80 | 50.86 | 26.5 | 6.5 | 35.65 | 46.62 |
| VAE [42] | 72.02 | 46.07 | 25.4 | 7.9 | 39.33 | 53.36‡ | **75.53** | 42.83 | 25.8 | 6.2 | 37.95 | 52.74‡ |
| PLACE [43] | 64.81 | 57.26 | 25.1 | 8.2 | 42.64 | 54.24‡ | 62.49 | 57.34 | 26.0 | 5.3 | 41.44 | 54.00‡ |
| RAM-DSIR [44] | 54.42 | 61.55 | 23.6 | 6.7 | 40.30 | 51.98‡ | 61.25 | 57.71 | 24.3 | 5.4 | 41.44 | 54.03 ‡ |
| AdaptSeg [18] | 60.91 | 72.74 | 19.2 | 3.8 | 48.54 | 60.92‡ | 62.28 | 58.07 | 24.8 | 4.1 | 41.51 | 54.89‡ |
| PnP-AdaNet [45] | 65.92 | 69.59 | 20.6 | 3.0 | 48.93 | 61.55‡ | 62.30 | 57.41 | 19.2 | 3.8 | 42.39 | 55.72‡ |
| Advent [46] | 65.61 | 69.92 | 17.9 | 3.1 | 50.06 | 62.98‡ | 64.43 | 55.97 | 19.7 | 3.8 | 42.62 | 56.26‡ |
| AdaptPatch [47] | 53.38 | 74.75 | 16.2 | 4.3 | 43.94 | 54.84‡ | 66.63 | 56.24 | 25.7 | 4.3 | 44.28 | 57.16‡ |
| DDFseg [48] | 73.11 | 73.34 | 13.8 | 2.1 | 57.50 | 70.73‡ | 62.10 | 59.39 | 17.9 | 3.4 | 44.11 | 58.15‡ |
| SynSeg-Net [13] | 64.61 | **82.16** | 11.1 | 1.9 | 53.19 | 63.32‡ | 62.96 | **74.80** | 14.9 | 3.1 | 50.96 | 62.33‡ |
| CyCADA [33] | 70.83 | 77.48 | 14.5 | 2.1 | 56.50 | 67.77‡ | 68.04 | 70.17 | 18.9 | 3.0 | 51.81 | 63.99‡ |
| CycleGAN [12] | 73.24 | 69.15 | 16.4 | 3.3 | 53.77 | 66.63‡ | 66.03 | 71.28 | 19.8 | 3.4 | 51.33 | 64.11‡ |
| Prior SIFA [38] | 74.61 | 70.01 | 13.0 | 1.8 | 56.38 | 70.32‡ | 66.56 | 67.04 | 16.4 | 2.7 | 50.12 | 64.19‡ |
| UESM [40] | 75.58 | 74.35 | 14.4 | 1.8 | 59.93 | 73.11‡ | 71.69 | 63.75 | 24.0 | 2.6 | 50.78 | 64.97‡ |
| DSAN [39] | 78.45 | 72.57 | 12.2 | 1.8 | 60.26 | 73.45‡ | 69.95 | 66.26 | 15.3 | 2.5 | 51.69 | 65.71‡ |
| SASAN [31] | 74.64 | 68.01 | 25.9 | 2.2 | 53.62 | 67.05‡ | 72.49 | 67.18 | 20.8 | 3.0 | 52.92 | 66.37‡ |
| SIFA [21] | 72.92 | 72.68 | 11.7 | 2.1 | 57.07 | 70.45‡ | 68.95 | 68.90 | 14.4 | 2.4 | 52.50 | 66.42‡ |
| SAM [49] | 76.33 | 73.51 | 15.7 | 4.4 | 59.12 | 71.96‡ | 73.52 | 68.48 | 23.1 | 6.3 | 54.15 | 67.68‡ |
| Ours | **80.61** | 77.66 | **10.8** | **1.4** | **64.82** | **76.52** | 73.02 | 74.53 | **13.9** | **2.3** | **57.87** | **70.04** |

‡ denotes $p < 0.001$ of two-tailed Wilcoxon test.

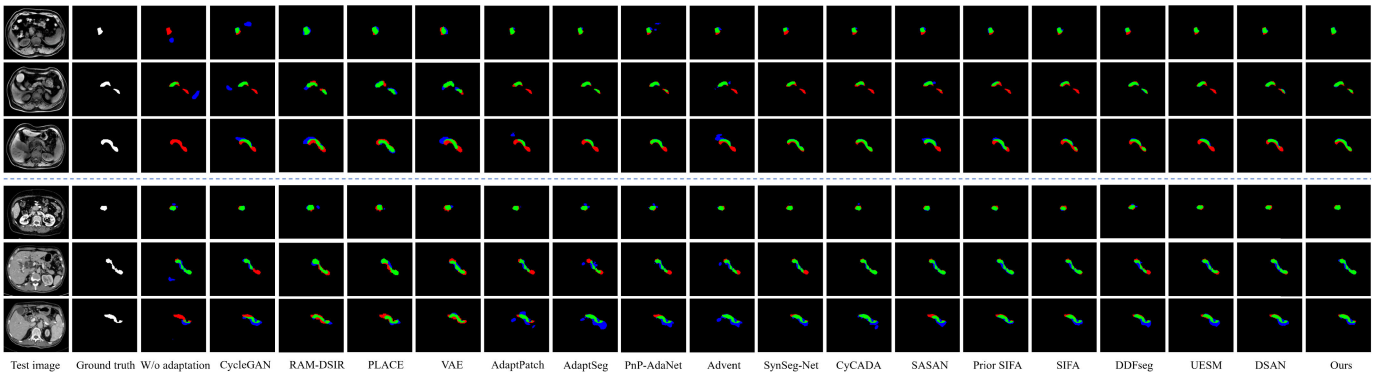


Fig. 5. Visual comparison of segmented results produced by different methods for pancreas MR images (top three rows) and CT images (bottom three rows). From left to right are the raw test images (1st column), ground truth (2nd column), "w/o Adaptation" lower bound (3nd column), results of other unsupervised domain adaptation methods (4rd-19th column) and results of our method (last column).

was shown in Fig.5. Our method was capable of accurately segmenting the pancreas, while other methods tended to produce incorrect segmentation.

As shown in Table III, for CT to MR, the Recall score and the Precision score of VAE were respectively 72.02 and 46.07; and for MR to CT, the Recall score and the Precision score were respectively 75.53 and 42.83. The high Recall scores and low Precision scores of CT to MR and MR to CT showed that VAE produced severe over-segmentation and many non-pancreatic tissues were incorrectly segmented as the pancreas, because the VAE-based framework maybe failed to drive two domains to one parameterized distribution with a sliced distance. PLACE utilized a randomly source-domain feature-level augmentation strategy without learning the characteristics of the target domain, and RAM-DSIR introduced a random amplitude mixup module by utilizing the Fourier transform to augment source domain images. These methods did not synthetize target domain image and use the corresponding ground truth, and therefore, severe under-segmentation and over-segmentation were both led because of the large domain shift. AdaptSeg adopted adversarial learning to make target predictions closer to the source ones, considering semantic segmentations as structured outputs that contain spatial similarities between the source and target. PnP-AdaNet computed Wassertein distance between the multiple-level features of the source domain and the target domain and also computed Wassertein distance between the segmentation predictions for the target and source domains. DDFseg, DSAN, SIFA and Prior SIFA also thought the anatomical shapes were similar between the source and target domains, and they also directly

TABLE IV
PERFORMANCE COMPARISON WITH DIFFERENT UNSUPERVISED DOMAIN ADAPTATION METHODS FOR AMOS

| Method | CT→MR | | | | | | MR→CT | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Recall (%) | Precision (%) | $HD_{95}$ (mm) | ASD (mm) | Jac (%) | DSC (%) | Recall (%) | Precision (%) | $HD_{95}$ (mm) | ASD (mm) | Jac (%) | DSC (%)↑ |
| Supervised training | 85.96 | 84.99 | 8.3 | 1.8 | 71.46 | 82.08 | 80.99 | 82.11 | 10.4 | 2.0 | 68.42 | 79.37 |
| w/o adaptation | 36.06 | 77.36 | 17.6 | 6.0 | 30.44 | 38.11 | 47.43 | 50.89 | 28.5 | 10.3 | 31.75 | 40.97 |
| SASAN [31] | 49.79 | 60.10 | 29.3 | 7.8 | 36.57 | 50.01‡ | 53.48 | 44.49 | 27.5 | 9.8 | 31.10 | 43.14‡ |
| PnP-AdaNet [45] | 61.10 | 68.32 | 19.9 | 5.5 | 44.64 | 56.79‡ | 50.67 | 54.01 | 26.8 | 9.0 | 30.60 | 41.46‡ |
| RAM-DSIR [44] | 60.13 | 78.67 | 16.1 | 3.8 | 51.02 | 62.77‡ | 50.0 | 61.78 | 23.5 | 8.0 | 35.66 | 45.44‡ |
| DDFseg [48] | 71.13 | 61.59 | 16.5 | 4.5 | 49.19 | 62.81‡ | 52.02 | 51.17 | 21.5 | 7.5 | 36.14 | 48.44‡ |
| PLACE [43] | 74.89 | 64.13 | 27.7 | 6.1 | 50.75 | 63.03‡ | 57.65 | 53.25 | 24.4 | 8.1 | 38.08 | 48.73‡ |
| AdaptPatch [47] | 68.87 | 74.38 | 15.9 | 3.8 | 54.84 | 67.10‡ | 63.94 | 48.77 | 25.5 | 7.4 | 37.89 | 49.58‡ |
| VAE [42] | 76.97 | 56.73 | 26.3 | 5.8 | 47.46 | 61.77‡ | 73.26 | 44.97 | 23.7 | 6.0 | 37.72 | 52.22‡ |
| AdaptSeg [18] | 65.45 | 66.14 | 18.8 | 4.9 | 48.51 | 60.78‡ | 61.99 | 52.06 | 22.5 | 6.7 | 38.95 | 50.58‡ |
| CyCADA [33] | 72.81 | 67.47 | 17.3 | 4.0 | 52.43 | 65.33‡ | 61.32 | 59.53 | 21.6 | 6.3 | 43.10 | 55.10‡ |
| CycleGAN [12] | 73.44 | 75.48 | 14.1 | 3.3 | 58.60 | 70.54‡ | 57.81 | **69.67** | 18.2 | 5.2 | 45.08 | 56.66‡ |
| Prior SIFA [38] | 78.56 | 70.53 | 12.1 | 3.1 | 59.15 | 72.50‡ | 64.32 | 55.70 | 18.4 | 5.6 | 43.29 | 56.79‡ |
| UESM [40] | 75.56 | 77.77 | 11.8 | 2.7 | 61.37 | 73.58‡ | 62.77 | 63.09 | 19.4 | 5.4 | 45.14 | 56.82‡ |
| DSAN [39] | 74.11 | 67.32 | 14.3 | 3.9 | 54.73 | 68.10‡ | 66.66 | 56.22 | 17.0 | 5.2 | 44.97 | 58.63‡ |
| SIFA [21] | 76.27 | 72.26 | 12.0 | 3.1 | 59.09 | 72.12‡ | 66.98 | 59.56 | 16.0 | 4.7 | 46.20 | 59.63‡ |
| Advent [46] | 72.08 | 66.80 | 18.1 | 4.4 | 51.27 | 63.81‡ | 73.27 | 58.73 | 18.4 | 4.9 | 47.69 | 60.40‡ |
| SynSeg-Net [13] | 78.98 | 75.40 | 12.4 | 2.7 | 61.91 | 74.16‡ | 68.54 | 66.59 | 16.1 | 4.3 | 49.58 | 61.71† |
| SAM [49] | 79.96 | 64.69 | **9.3** | 2.8 | 54.39 | 68.77‡ | **77.44** | 64.11 | **9.5** | **2.6** | **53.29** | **67.23**‡ |
| Ours | **81.10** | **80.69** | 9.9 | **2.0** | **67.26** | **78.94** | 67.99 | 68.46 | 17.0 | 4.4 | 50.68 | 62.60 |

‡ denotes $p < 0.001$ of two-tailed Wilcoxon test, † denotes $p < 0.05$ of two-tailed Wilcoxon test.

aligned their segmented results of the target and source domain images though the unpaired images had different ground truth. In our study, the anatomy of the pancreas varied largely between unpaired the target and source domain images, the shapes of the pancreas should not be forced to be the similar, as shown in Fig.5. Advent introduced the Shannon entropy loss to directly maximize prediction certainty in the target domain and used a simple class-prior based on the distribution of the classes over the source labels and imposed the class prior constraint on the target image. However, the shape and histogram of number of pixels per class may be largely different between the unpaired source domain images and target domain images. As shown in Fig.5, Advent tended to incorrectly segment the pancreas with complex shape. UESM combined the uncertainty-aware self-training scheme with an adversarial learning block to align the extracted features of source and target domains. Although they proposed an uncertainty estimation and segmentation module to obtain the uncertainty map estimation of the target domain, they used the pseudo labels of target domain images, and did not explicitly use the source label to improve the segmentation of target domain images. SASAN added another attention branch in PatchGAN to extract orthogonal attention features which was used to transfer the style, but pseudo labels derived from the attention module was simply used to train the segmentation network. CyCADA pretrained source task model and performed image synthesis with CycleGAN. SynSeg-Net proposed an end-to-end synthetic segmentation network with CycleGAN. Although the Precision scores of CT to MR and MR to CT were high but the Recall scores were relatively low, which showed that severe under-segmentation was led. Due to the success of GAN, CycleGAN-based methods allowed to transform source domain images into target-like domain images, and target-like domain images with ground truth of the corresponding source domain images can be used to train a target domain image segmentation network. However, CyCADA, SynSeg-Net and CycleGAN did not pay attention to the shape consistency in the process of image synthesis or generation. As indicated by qualitative results, the proposed feature contrastive loss term and the moment consistency loss ensured preserving organs or tissues shape-invariant, thus facilitating the translation process across image modalities.

### *F. Comparison With State-of-the-Art Methods in AMOS Dataset*

The results of unsupervised domain adaption (CT→MR) and (MR→CT) methods for the AMOS dataset were shown in Table IV. Similarly, 'w/o adaption' only obtained 36.06(Recall), 77.36(Precision), 17.6($HD_{95}$), 6.0(ASD), 30.44(Jac) and 38.11(DSC) in the CT to MR domain adaption and 47.43(Recall), 50.89(Precision), 28.5($HD_{95}$), 10.3(ASD), 31.75(Jac) and 40.97(DSC) in the MR to CT domain adaption, showing significant performance degradation compared to supervised training. Compared to MMPS dataset, all scores of SASAN dramatically decreased in AMOS dataset because pseudo labels derived from the attention module probably led to negative effect due to the complexity of AMOS dataset. In the MR to CT domain adaption, the Jac and DSC scores of PnP-AdaNet were only 30.60 and 41.46, respectively. Due to the limited number of MR images, it was difficult to minimize Wassertein distance between the multiple-level features of the

TABLE V
PERFORMANCE COMPARISON WITH DIFFERENT UNSUPERVISED DOMAIN ADAPTATION METHODS FOR MMWHS

| Method | CT→MR | | | | | | | | | | MR→CT | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | DSC(%) | | | | | ASD(mm) | | | | | DSC(%) | | | | | ASD(mm) | | | | |
| | AA | LAC | LVC | MYO | Avg | AA | LAC | LVC | MYO | Avg | AA | LAC | LVC | MYO | Avg↑ | AA | LAC | LVC | MYO | Avg |
| Supervised training | 82.8 | 80.5 | 92.4 | 78.8 | 83.6 | 3.6 | 3.9 | 2.1 | 1.9 | 2.9 | 92.7 | 91.1 | 91.9 | 87.7 | 90.9 | 1.5 | 3.5 | 1.7 | 2.1 | 2.2 |
| w/o adaptation | 5.4 | 30.2 | 24.6 | 2.7 | 15.7 | 15.4 | 16.8 | 13.0 | 10.8 | 14.0 | 28.4 | 27.7 | 4.0 | 8.7 | 17.2 | 20.6 | 16.2 | N/A | 48.4 | N/A |
| CycleGAN [12] | 64.3 | 30.7 | 65.0 | 43.0 | 50.7 | 5.8 | 9.8 | 6.0 | 5.0 | 6.6 | 73.8 | 75.7 | 52.3 | 28.7 | 57.6 | 11.5 | 13.6 | 9.2 | 8.8 | 10.8 |
| SynSeg-Net [13] | 41.3 | 57.5 | 63.6 | 36.5 | 49.7 | 8.6 | 10.7 | 5.4 | 5.9 | 7.6 | 71.6 | 69.0 | 51.6 | 40.8 | 58.2 | 11.7 | 7.8 | 7.0 | 9.2 | 8.9 |
| AdaptSeg [18] | 60.8 | 39.8 | 71.5 | 35.5 | 51.9 | 5.7 | 8.0 | 4.6 | 4.6 | 5.7 | 65.2 | 76.6 | 54.4 | 43.6 | 59.9 | 17.9 | 5.5 | 5.9 | 8.9 | 9.6 |
| VAE [42] | 53.4 | 67.7 | 72.5 | 66.3 | 65.0 | 15.7 | 11.2 | 11.7 | 8.6 | 11.8 | 55.8 | 67.8 | 77.8 | 53.3 | 63.7 | 12.7 | 11.6 | 8.3 | 9.2 | 10.5 |
| PnP-AdaNet [45] | 43.7 | 47.0 | 77.7 | 48.6 | 54.3 | 11.4 | 14.5 | 4.5 | 5.3 | 8.9 | 74.0 | 68.9 | 61.9 | 50.8 | 63.9 | 12.8 | 6.3 | 17.4 | 14.7 | 12.8 |
| CyCADA [33] | 60.5 | 44.0 | 77.6 | 47.9 | 57.5 | 7.7 | 13.9 | 4.8 | 5.2 | 7.9 | 72.9 | 77.0 | 62.4 | 45.3 | 64.4 | 9.6 | 8.0 | 9.6 | 10.5 | 9.4 |
| Prior SIFA [38] | 67.0 | 60.7 | 75.1 | 45.8 | 62.1 | 6.2 | 9.8 | 4.4 | 4.4 | 6.2 | 81.1 | 76.4 | 75.7 | 58.7 | 73.0 | 10.6 | 7.4 | 6.7 | 7.8 | 8.1 |
| RAM-DSIR [44] | 50.0 | 61.5 | 71.2 | 50.6 | 58.3 | 21.8 | 12.1 | 7.3 | 9.3 | 12.6 | 70.1 | 74.6 | 82.0 | 65.8 | 73.1 | 13.5 | 10.2 | 7.2 | 7.4 | 9.6 |
| SIFA [21] | 65.3 | 62.3 | 78.9 | 47.3 | 63.4 | 7.3 | 7.4 | 3.8 | 4.4 | 5.7 | 81.3 | 79.5 | 73.8 | 61.6 | 74.1 | 7.9 | 6.2 | 5.5 | 8.5 | 7.0 |
| PLACE [43] | 48.0 | 72.0 | 75.2 | 60.0 | 63.8 | 11.0 | 10.5 | 12.7 | 9.0 | 10.8 | 71.2 | 76.1 | 83.8 | 67.5 | 74.6 | 13.0 | 9.3 | 5.9 | 7.8 | 9.0 |
| SAM [49] | **76.7** | **77.6** | 80.8 | 56.5 | **72.9** | 5.9 | 6.4 | 5.4 | 7.2 | 6.2 | **87.4** | 83.6 | 85.7 | 52.7 | 77.4 | **2.5** | **4.4** | 4.1 | 7.6 | 4.7 |
| SASAN [31] | 54.4 | 73.4 | **86.6** | 68.1 | 70.6 | 18.8 | 9.4 | 6.1 | 3.9 | 9.5 | 82.1 | 76.3 | 82.4 | 72.7 | 78.4 | 4.1 | 8.3 | 3.5 | **3.3** | 4.9 |
| DSAN [39] | 71.3 | 66.2 | 76.2 | 52.1 | 66.5 | **4.4** | 7.3 | 5.5 | 4.3 | 5.4 | 79.9 | 84.8 | 82.8 | 66.5 | 78.5 | 7.7 | 6.7 | 3.8 | 5.6 | 5.9 |
| UESM [40] | 69.5 | 67.0 | 75.4 | 60.2 | 68.0 | 4.9 | 4.5 | **3.2** | 4.6 | 4.3 | 82.1 | **87.0** | 83.6 | 68.4 | 80.3 | 4.2 | 5.3 | 4.1 | 4.5 | 4.5 |
| Ours | 58.9 | 74.8 | 84.3 | **69.2** | 71.8 | 5.4 | 4.1 | 3.9 | **3.5** | **4.2** | 75.2 | 82.6 | **89.6** | **74.4** | **80.5** | 4.1 | 4.9 | **3.1** | 4.8 | **4.2** |

TABLE VI
PERFORMANCE COMPARISON USING DIFFERENT NUMBERS OF TEACHER NETWORKS ON MMPS

| Number | CT→MR | | | | | | MR→CT | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Recall (%) | Precision (%) | $HD_{95}$ (mm) | ASD (mm) | Jac (%) | DSC (%) | Recall (%) | Precision (%) | $HD_{95}$ (mm) | ASD (mm) | Jac (%) | DSC (%) |
| 1 | 77.58 | 77.82 | 12.6 | 3.4 | 62.68 | 74.06 | **82.53** | 62.78 | 24.0 | 5.1 | 54.53 | 68.16 |
| 2 | 80.49 | 74.80 | 13.7 | 3.4 | 62.43 | 74.55 | 78.82 | 67.71 | 22.6 | 4.9 | 56.45 | 69.07 |
| 4 | **80.61** | 77.66 | **10.8** | **1.4** | **64.82** | **76.52** | 73.02 | **74.53** | **13.9** | **2.3** | **57.87** | **70.04** |
| 6 | 80.46 | **78.24** | 12.2 | 3.1 | 64.81 | 76.12 | 78.16 | 68.97 | 18.3 | 4.2 | 56.50 | 69.76 |

TABLE VII
PERFORMANCE OF SAM WITH DIFFERENT PROMPTS FOR MMPS

| Prompt | MR | | | | | | CT | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Recall (%) | Precision (%) | $HD_{95}$ (mm) | ASD (mm) | Jac (%) | DSC (%) | Recall (%) | Precision (%) | $HD_{95}$ (mm) | ASD (mm) | Jac (%) | DSC (%) |
| P0 | 79.88 | 4.58 | 165.58 | 48.4 | 2.63 | 5.05 | 65.98 | 3.62 | 211.19 | 52.33 | 3.52 | 6.7 |
| P1 | 90.80 | 38.41 | 85.2 | 19.3 | 36.02 | 49.82 | 86.75 | 46.48 | 85.4 | 18.7 | 41.57 | 55.52 |
| P2 | **94.21** | 56.29 | 49.1 | 9.9 | 53.69 | 67.01 | **96.60** | 45.39 | 77.4 | 16.9 | 44.26 | 58.08 |
| P3 | 72.12 | 71.61 | **15.1** | 4.5 | 56.10 | 68.66 | 67.74 | 66.71 | **21.8** | 6.4 | 51.38 | 64.24 |
| P4 | 76.33 | **73.51** | 15.7 | **4.4** | **59.12** | **71.96** | 73.52 | **68.48** | 23.1 | **6.3** | **54.15** | **67.68** |

source domain and the target domain. To be similar to MMPS dataset, the high Recall scores and low Precision scores of CT to MR and MR to CT of AMOS dataset also showed that VAE produced severe over-segmentation and many non-pancreatic tissues were incorrectly segmented. Similarly, Advent also produced severe over-segmentation, which also showed that the class prior constraint of the source labels may not be consistent with the target image due to the wide variations in shape on the different patients. CyCADA and CycleGAN still held the moderate performance on AMOS dataset since they did not maintain the shape consistency in the process of image synthesis. For SynSeg-Net and our method, he Recall scores and Precision scores of CT to MR and MR to CT were both balanced. For CT to MR domain adaption, our proposed method outperformed all the state-of-the-art methods in all metrics. The Recall, Precision, Jac and DSC scores improved to 81.10, 80.69, 67.26 and 78.94, respectively, and the ASD and $HD_{95}$ decreased to 2.0 and 9.9, respectively. For MR to CT domain adaption, our method also achieved the best performance in terms of Jac and DSC scores, which reached 50.68(Jac) and 62.60(DSC). In the comparison of other metrics, our method was very close to the top-performing methods. The experiment results on the AMOS dataset also proved the superiority of our method. $p < 0.05$ of the DSC score showed that the superiority of our method for pancreas segmentation was statistically significant.

### G. Comparison With State-of-the-Art Methods in MMWHS Dataset

Table V shows the segmented results of unsupervised domain adaptation (CT→MR) and (MR→CT) methods for the cardiac dataset, where cross-modality performance degradation was similar to the pancreas dataset. Without domain adaptation, the model trained on CT images and directly tested on MR images obtained 15.7 in terms of the average DSC score, and the average ASD was 14.0. The model trained on MR images and directly tested on CT images obtained 17.2 in terms of the average DSC score, and the average ASD was very large(N/A). For CT→MR domain adaptation, there existed a large performance gap (67.9 in terms of the average DSC score, 11.1 in terms of ASD) to the supervised training. For MR→CT domain adaptation, there also existed a large performance gap (73.7 in terms of the average DSC score) to the supervised training. Although CycleGAN holded the moderate performance and SynSeg-Net performed well on AMOS dataset, all scores of CycleGAN and SynSeg-Net dramatically decreased in MMWHS dataset due to the complexity of cardiac structures. PnP-AdaNet also obtained a low average DSC score of CT to MR, which also showed that it was probably infeasible to directly aligned the multiple-level features and segmented results of the target and source domain images since the unpaired images had different shapes. For MR to CT, the average DSC score of RAM-DSIR improved to 73.1 and the average DSC score of PLACE improved

TABLE VIII
NOISE PERTURBATION EXPERIMENTS FOR CT->MR AND MT->CT DOMAIN ADAPTATION FOR MMPS

| Method | CT→MR | | | | | | MR→CT | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Recall (%) | Precision (%) | $HD_{95}$ (mm) | ASD (mm) | Jac (%) | DSC (%) | Recall (%) | Precision (%) | $HD_{95}$ (mm) | ASD (mm) | Jac (%) | DSC (%) |
| DDFseg | 73.11 | 73.34 | 13.8 | 2.1 | 57.50 | 70.73 | 62.10 | 59.39 | 17.9 | 3.4 | 44.11 | 58.15 |
| DDFseg($\sigma_1$) | 70.62 | 69.43 | 14.7 | 4.0 | 53.94 | 67.71 | 65.13 | 55.91 | 18.4 | 5.2 | 42.80 | 57.31 |
| DDFseg($\sigma_2$) | 70.47 | 68.22 | 15.0 | 4.5 | 53.49 | 66.72 | 62.86 | 55.17 | 17.9 | 5.2 | 41.31 | 55.82 |
| DDFseg($\sigma_3$) | 69.81 | 66.90 | 16.4 | 4.9 | 52.56 | 65.62 | 59.07 | 54.29 | 18.4 | 5.6 | 39.29 | 53.53 |
| UESM | 75.58 | 74.35 | 14.4 | 1.8 | 59.93 | 73.11 | 71.69 | 63.75 | 24.0 | 2.6 | 50.78 | 64.97 |
| UESM($\sigma_1$) | 82.40 | 69.59 | 15.9 | 3.7 | 59.25 | 72.50 | 73.89 | 65.15 | 26.8 | 5.9 | 49.45 | 63.52 |
| UESM($\sigma_2$) | 82.14 | 65.82 | 17.2 | 4.2 | 56.55 | 70.03 | 77.29 | 56.56 | 43.4 | 7.9 | 46.47 | 61.21 |
| UESM($\sigma_3$) | 81.07 | 65.41 | 17.0 | 4.2 | 54.16 | 68.09 | 70.79 | 60.10 | 50.0 | 8.0 | 44.91 | 59.12 |
| DSAN | 78.45 | 72.57 | 12.2 | 1.8 | 60.26 | 73.45 | 69.95 | 66.26 | 15.3 | 2.5 | 51.69 | 65.71 |
| DSAN($\sigma_1$) | 78.54 | 69.18 | 13.3 | 3.6 | 58.40 | 71.71 | 71.10 | 64.41 | 14.7 | 4.4 | 51.00 | 65.14 |
| DSAN($\sigma_2$) | 79.61 | 67.28 | 14.7 | 3.8 | 57.71 | 71.11 | 70.37 | 60.76 | 15.7 | 4.6 | 48.55 | 62.81 |
| DSAN($\sigma_3$) | 77.99 | 60.14 | 16.8 | 4.7 | 51.96 | 65.99 | 67.87 | 54.53 | 18.1 | 5.5 | 43.24 | 57.86 |
| Ours | 80.61 | 77.66 | 10.8 | 1.4 | 64.82 | 76.52 | 73.02 | 74.53 | 13.9 | 2.3 | 57.87 | 70.04 |
| Ours($\sigma_1$) | 78.56 | 79.53 | 11.5 | 3.0 | 64.41 | 75.51 | 76.02 | 68.66 | 16.9 | 4.5 | 56.07 | 68.87 |
| Ours($\sigma_2$) | 76.44 | 76.44 | 14.1 | 3.4 | 61.28 | 72.91 | 72.42 | 71.17 | 15.2 | 3.9 | 55.02 | 67.73 |
| Ours($\sigma_3$) | 75.25 | 76.22 | 12.7 | 3.2 | 59.79 | 71.63 | 69.80 | 70.61 | 16.2 | 4.5 | 53.70 | 66.14 |

to 74.6, which showed that the randomly source-domain augmentation strategy might be beneficial to local contrast-enhanced CT images of cardiac structures, but they obtained moderate DSC scores due to the relatively low contrast of MR images. Our method outperformed state-of-the-art approaches. For CT→MR domain adaptation, the average DSC score improved to 71.8 and the average ASD decreased to 4.2 for the cardiac segmentation. For MR→CT domain adaptation, the average DSC score improved to 80.5 and the average ASD also decreased to 4.2 for the cardiac segmentation.

### H. Analysis of SAM

As shown in Table III, Table IV and Table V, despite not being trained on medical images, SAM still achieved comparable performance. On MMPS dataset, our method outperformed SAM. On AMOS dataset, our method surpassed SAM by 10.14(DSC) in CT→MR domain adaption but decreased by 4.63(DSC) in MR→CT domain adaption due to extreme domain imbalance. As for MMWHS dataset, SAM slightly outperformed our method in CT→MR domain adaption in terms of DSC, but reached larger ASD than our method in CT→MR domain adaption. Our method surpassed SAM by 3.1(DSC) and reached smaller ASD in MR→CT domain adaption.

SAM was a semi-automatic segmentation method that required manual input prompts for image segmentation. ViT-H was used as the backbone. For the prompts, we deployed five groups: P0 was no input prompt, P1 was one positive point, P2 was five positive and five negative points, P3 was one box, and P4 was one box and one positive point. Following [50], the bounding box prompt was simulated from the expert annotations with a random perturbation of 0-20 pixels. Positive points were selected in the center of the pancreas, and negative points were randomly selected around the pancreas. First, it heavily relied on manual input prompts, and the accuracy of these prompts greatly impacted the final segmented results. It still required significant human involvement. As shown in Table VII, without providing prompts, the segmented results of SAM were consistently poor across two modalities. In contrast, our method could achieve fully automatic pancreatic segmentation, which held greater practical significance.

Another disadvantage of SAM was its inability to correctly discern structural relationships. The experiments on the three datasets show that in most cases, SAM reached high Recall scores but low Precision scores, i.e., significant over-segmentation was produced. SAM tended to segment all possible regions as foreground based on the prompts.

### I. Perturbation Experiments With Gaussian Noise

Several noisy perturbation experiments on MMPS dataset have been performed to evaluate the robustness of our method. Three comparison methods with good performance were also tested with noisy perturbation experiments. Table VIII showed the results of perturbation experiments with three levels Gaussian noise, i.e. three variances $\sigma_1, \sigma_2, \sigma_3$ of Gaussian noises were 0.001, 0.005 and 0.01, respectively. Our proposed method decreased 1.01 and 0.41 in the CT→MR adaption and 1.17 and 1.80 in the MR→CT adaption for DSC and Jac scores when the $\sigma$ of Gaussian noise was 0.001, whereas DDFseg, UESM and DSAN decreased by 3.02(DSC) and 3.56(Jac), 0.61(DSC) and 0.68(Jac), 1.74(DSC) and 1.86(Jac) in the CT→MR adaption, and 0.84(DSC) and 1.31(Jac), 1.45(DSC) and 1.33(Jac), 0.57(DSC) and 0.69(Jac) in the MR→CT adaptation, respectively. At lower values of $\sigma$, all methods only showed minor performance degradation. However, as $\sigma$ increased, the performance loss of each method gradually became larger. This indicated that higher intensities of noise could deteriorate image quality and significantly impact model performance. For CT→MR adaption, all the methods showed similar levels of degradation. However, for MR→CT adaption, our method showed better stability compared to other methods. When $\sigma$ came to 0.01, DDFseg, UESM and DSAN led to decrease by 4.62(DSC), 5.85(DSC), and 7.85(DSC), while our method only decreased by 3.90(DSC).

## V. Discussion and Conclusion

In this paper, MC-CCycleGAN is proposed for cross-domain pancreas segmentation. The structure and shape of the pancreas are well preserved through our method, which significantly improves the segmentation performance. The results presented in the previous sections have demonstrated the effectiveness and superiority of our method.

Kindly note that the proposed framework for cross-domain pancreas segmentation is a kind of unsupervised domain adaptation, in which the source domain images are provided with segmentation labels but the target domain images are not provided with segmentation labels in the training stage. Although the scenario is called unsupervised domain adaptation, it is indeed a kind of semi-supervised learning, which needs the labels of source domain images for cross-domain adaptation.

Unsupervised domain adaptation aims to utilize labeled source domain images to adapt to the target domain without using additional annotations from the target domain. CycleGAN is often employed for unsupervised domain adaptation as it allows for transferring the style of unlabeled target domain images to labeled source domain images. The synthesized target domain images, along with the corresponding labels of the labeled source domain images, can be used to train the segmentation network, which can then be employed to segment the target domain images. Compared to a segmentation network trained solely with labeled source domain images, a segmentation network trained with synthesized target domain images can significantly improve the performance of target domain image segmentation. However, CycleGAN-based unsupervised domain adaptation methods tend to globally transfer the style of unlabeled target domain images to labeled source domain images, without focusing on the structural features of the input image and the shape of the object.

Our CCycleGAN aligns the structural features of the synthesized images with the original images at the feature level. A contrastive loss has been designed to explicitly ensure the full extraction of structural features from source images while remove redundant style features in the process of style transfer. On the one hand, the positive samples are consistent with the query samples in structure and shape but different in styles. The encoder $E$ in the generator $G_t$ is applied to the positive samples $x^{s\to t}$ (synthesized target domain images), so as to extract the features of the structure and content, to capture domain-invariant features and to remove the style features. The negative samples are not consistent in structure and shape with the query samples although they have a similar style. On the other hand, contrastive learning can bring the query sample and the positive sample closer to each other but pull apart negative samples. By imposing the constraint of feature contrastive learning, the encoder $E$ tends to gradually remove the style at the feature level, capture domain-invariant features and keep the consistency of structure and shape in the training of style transfer.

In addition, to maintain the consistency of the structure and shape of the pancreas in both the original and synthesized images, shape-level alignment is also integrated into our method. Multi-order central moments are introduced and integrated into the contrastive loss to maintain consistency in the structure of the pancreas before and after style transfer. Multi-order central moments can abstractly describe the structure of the predicted pancreas and provide a high-dimensional description of its anatomy. When the structure of the segmented pancreas changes, its multi-order central moments change. This characteristic can be used to maintain the consistency of pancreatic shape before and after image transformation. Although multi-order central moments are used in our contrastive learning to evaluate moment consistency, they can be used to image registration and other image segmentation tasks.

With regard to multi-teacher knowledge distillation, four teacher networks are built to jointly guide the training of a single student network for better segmented results. Synthesized images are selected from four latter epochs of the image transformation subnetwork training to feed into the teacher and student networks. The segmentation network in the SynSeg-Net is trained together with the image transformation network and fed with all synthesized images from all epochs, which may result in the potential collapse of the entire segmentation network. Our selected images are very similar in structure while have subtle differences in style of the target domain images; and therefore, this can be served as data augmentation which can provide more images with different styles and improve the student network.

Overall, comprehensive experiments illustrate the superior performance of our proposed method in the cross-domain pancreas segmentation. Compared to state-of-the-art methods, our approach tends to preserve the structure of the pancreas. CT imaging can provide non-enhanced, venous phase and arterial phase images and MR imaging can also provide T1 and T2 images; and therefore, we will try more imaging modalities and collect larger number of multi-modality images to show the potential multi-domain adaptation of our proposed method.

## References

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2020," *CA: Cancer J. Clinicians*, vol. 70, no. 1, pp. 7–30, 2020.

[2] S. V. Shrikhande, S. G. Barreto, M. Goel, and S. Arya, "Multimodality imaging of pancreatic ductal adenocarcinoma: A review of the literature," *HPB*, vol. 14, no. 10, pp. 658–668, Oct. 2012.

[3] Z. Zhu, Y. Xia, W. Shen, E. Fishman, and A. Yuille, "A 3D coarse-to-fine framework for volumetric medical image segmentation," in *Proc. Int. Conf. 3D Vis. (3DV)*, Sep. 2018, pp. 682–690.

[4] Z. Zhou et al., "A dual branch and fine-grained enhancement network for pancreatic tumor segmentation in contrast enhanced CT images," *Biomed. Signal Process. Control*, vol. 82, Apr. 2023, Art. no. 104516.

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

[6] M. Ghafoorian et al., "Transfer learning for domain adaptation in MRI: Application in brain lesion segmentation," in *Proc. 20th Int. Conf. Med. Image Comput. Comput. Assist. Intervent. (MICCAI)*, Quebec City, QC, Canada. Cham, Switzerland: Springer, 2017, pp. 516–524.

[7] E. Gibson et al., "Inter-site variability in prostate segmentation accuracy using deep learning," in *Proc. 21st Int. Conf. Med. Image Comput. Comput. Assist. Intervent. (MICCAI)*, Granada, Spain. Cham, Switzerland: Springer, 2018, pp. 506–514.

[8] K. Kamnitsas et al., "Unsupervised domain adaptation in brain lesion segmentation with adversarial networks," in *Proc. 25th Int. Conf. Inf. Process. Med. Imag. (IPMI)*, Boone, NC, USA. Cham, Switzerland: Springer, 2017, pp. 597–609.

[9] 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," 2018, *arXiv:1804.10916*.

[10] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1125–1134.
[11] I. Goodfellow et al., "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.
[12] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2223–2232.
[13] Y. Huo et al., "SynSeg-net: Synthetic segmentation without target modality ground truth," *IEEE Trans. Med. Imag.*, vol. 38, no. 4, pp. 1016–1025, Apr. 2019.
[14] Q. Dou et al., "PnP-AdaNet: Plug-and-play adversarial domain adaptation network at unpaired cross-modality cardiac segmentation," *IEEE Access*, vol. 7, pp. 99065–99076, 2019.
[15] Q. Yu, L. Xie, Y. Wang, Y. Zhou, E. K. Fishman, and A. L. Yuille, "Recurrent saliency transformation network: Incorporating multi-stage visual cues for small organ segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 8280–8289.
[16] W. Xu et al., "Semi-supervised interactive fusion network for MR image segmentation," *Med. Phys.*, vol. 50, no. 3, pp. 1586–1600, Mar. 2023.
[17] A. Mathur, A. Isopoussu, F. Kawsar, N. B. Berthouze, and N. D. Lane, "FlexAdapt: Flexible cycle-consistent adversarial domain adaptation," in *Proc. 18th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2019, pp. 896–901.
[18] Y. Tsai, W. Hung, S. Schulter, K. Sohn, M. 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.
[19] V. Cheplygina, M. de Bruijne, and J. P. W. Pluim, "Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis," *Med. Image Anal.*, vol. 54, pp. 280–296, May 2019.
[20] 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.
[21] 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.
[22] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7167–7176.
[23] F. Wu and X. Zhuang, "CF distance: A new domain discrepancy metric and application to explicit domain adaptation for cross-modality cardiac image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 12, pp. 4274–4285, Dec. 2020.
[24] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 1–9.
[25] W.-L. Chang, H.-P. Wang, W.-H. Peng, and W.-C. Chiu, "All about structure: Adapting structural information across domains for boosting semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1900–1909.
[26] Q. Xie et al., "Unsupervised domain adaptation for medical image segmentation by disentanglement learning and self-training," *IEEE Trans. Med. Imag.*, vol. 43, no. 1, pp. 4–14, Jan. 2024.
[27] L. Bi, J. Kim, A. Kumar, D. Feng, and M. Fulham, "Synthesis of positron emission tomography (PET) images via multi-channel generative adversarial networks (GANs)," in *Proc. Mol. Imag., Reconstruction Anal. Moving Body Organs, Stroke Imag. Treatment: 5th Int. Workshop, CMMI, 2nd Int. Workshop, RAMBO, 1st Int. Workshop, SWITCH, Held Conjunct (MICCAI)*, Quebec City, QC, Canada. Cham, Switzerland: Springer, 2017, pp. 43–51.
[28] S. Dar, M. Yurt, L. Karacan, A. Erdem, E. Erdem, and T. Çukur, "Image synthesis in multi-contrast MRI with conditional generative adversarial networks," *IEEE Trans. Med. Imag.*, vol. 38, no. 10, pp. 2375–2388, Oct. 2019.
[29] J. Jiang et al., "Tumor-aware, adversarial domain adaptation from CT to MRI for lung cancer segmentation," in *Proc. 21st Int. Conf. Med. Image Comput. Comput. Assist. Intervent. (MICCAI)*, Granada, Spain. Cham, Switzerland: Springer, 2018, pp. 777–785.
[30] Y. Huo, Z. Xu, S. Bao, A. Assad, R. G. Abramson, and B. A. Landman, "Adversarial synthesis learning enables segmentation without target modality ground truth," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 1217–1220.
[31] D. Tomar, M. Lortkipanidze, G. Vray, B. Bozorgtabar, and J.-P. Thiran, "Self-attentive spatial adaptive normalization for cross-modality domain adaptation," *IEEE Trans. Med. Imag.*, vol. 40, no. 10, pp. 2926–2938, Oct. 2021.
[32] Y. Ganin et al., "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2030–2096, 2016.
[33] J. Hoffman et al., "CyCADA: Cycle-consistent adversarial domain adaptation," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1989–1998.
[34] Y. Zhang, Z. Qiu, T. Yao, D. Liu, and T. Mei, "Fully convolutional adaptation networks for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6810–6818.
[35] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.
[36] M.-K. Hu, "Visual pattern recognition by moment invariants," *IRE Trans. Inf. Theory*, vol. 8, no. 2, pp. 179–187, Feb. 1962.
[37] Y. Ji et al., "AMOS: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 36722–36732.
[38] C. Chen, Q. Dou, H. Chen, J. Qin, and P.-A. Heng, "Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 865–872.
[39] X. Han et al., "Deep symmetric adaptation network for cross-modality medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 41, no. 1, pp. 121–132, Jan. 2022.
[40] C. Bian et al., "Uncertainty-aware domain alignment for anatomical structure segmentation," *Med. Image Anal.*, vol. 64, Aug. 2020, Art. no. 101732.
[41] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*. Munich, Germany: Springer, 2015, pp. 234–241.
[42] F. Wu and X. Zhuang, "Unsupervised domain adaptation with variational approximation for cardiac segmentation," *IEEE Trans. Med. Imag.*, vol. 40, no. 12, pp. 3555–3567, Dec. 2021.
[43] J. Guo, L. Qi, Y. Shi, and Y. Gao, "PLACE dropout: A progressive layer-wise and channel-wise dropout for domain generalization," 2021, *arXiv:2112.03676*.
[44] Z. Zhou, L. Qi, and Y. Shi, "Generalizable medical image segmentation via random amplitude mixup and domain-specific image restoration," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 420–436.
[45] Q. Dou et al., "PnP-AdaNet: Plug-and-play adversarial domain adaptation network with a benchmark at cross-modality cardiac segmentation," 2018, *arXiv:1812.07907*.
[46] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, "Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 2517–2526.
[47] Y.-H. Tsai, K. Sohn, S. Schulter, and M. Chandraker, "Domain adaptation for structured output via discriminative patch representations," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1456–1465.
[48] C. Pei, F. Wu, L. Huang, and X. Zhuang, "Disentangle domain features for cross-modality cardiac image segmentation," *Med. Image Anal.*, vol. 71, Jul. 2021, Art. no. 102078.
[49] A. Kirillov et al., "Segment anything," 2023, *arXiv:2304.02643*.
[50] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," *Nature Commun.*, vol. 15, p. 654, Jan. 2024.