Medical Image Analysis 76 (2022) 102328

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

CyCMIS: Cycle-consistent Cross-domain Medical Image Segmentation via diverse image augmentation

Runze Wang[a,1], Guoyan Zheng[a,1,*]

Institute of Medical Robotics, School of Biomedical Engineering, Shanghai Jiao Tong University, No.800 Dongchuan Road, Shanghai 200240, China

ARTICLE INFO

Article history:
Received 6 July 2021
Revised 15 November 2021
Accepted 1 December 2021
Available online 8 December 2021

Keywords:
Cross-domain medical image segmentation
Diverse image translation
End-to-end
Unsupervised domain adaptation

ABSTRACT

Domain shift, a phenomenon when there exists distribution discrepancy between training dataset (source domain) and test dataset (target domain), is very common in practical applications and may cause significant performance degradation, which hinders the effective deployment of deep learning models to clinical settings. Adaptation algorithms to improve the model generalizability from source domain to target domain has significant practical value. In this paper, we investigate unsupervised domain adaptation (UDA) technique to train a cross-domain segmentation method which is robust to domain shift, and which does not require any annotations on the test domain. To this end, we propose Cycle-consistent Cross-domain Medical Image Segmentation, referred as CyCMIS, integrating online diverse image translation via disentangled representation learning and semantic consistency regularization into one network. Different from learning one-to-one mapping, our method characterizes the complex relationship between domains as many-to-many mapping. A novel diverse inter-domain semantic consistency loss is then proposed to regularize the cross-domain segmentation process. We additionally introduce an intra-domain semantic consistency loss to encourage the segmentation consistency between the original input and the image after cross-cycle reconstruction. We conduct comprehensive experiments on two publicly available datasets to evaluate the effectiveness of the proposed method. Results demonstrate the efficacy of the present approach.

1. Introduction

The past few years witnessed remarkable progress in medical image analysis due to the increasing availability of data and the rapid development of deep learning techniques (Litjens et al., 2017; Bernard et al., 2018; Zhuang et al., 2019; Wu et al., 2021b). Deep learning-based methods, such as convolutional neural networks (CNNs), are excellent at learning from large amounts of data (Yang et al., 2021; 2022), but can be poor at generalizing learned knowledge to new test datasets that differ from the training dataset (Choudhary et al., 2020; Guan and Liu, 2021). Such distribution discrepancy between the training dataset (source domain) and the test dataset (target domain), referred as domain shift, occurs frequently in medical imaging due to factors such as changing imaging modalities (e.g. computed tomography (CT), magnetic resonance imaging (MRI), Ultrasound, etc.), protocols (e.g. MRI protocols like PDw, T1w, T2w, etc.), scanning parameters (e.g. MRI pulse sequence parameters like repetition time, echo time, inversion time, flip angle, etc.), subject cohorts, or machines from different vendors and clinical sites. Domain shift is very common in practical applications and may cause significant performance degradation, which hinders the effective deployment of deep learning models to clinical settings. For semantic segmentation, which is a prerequisite for many clinical applications including disease diagnosis, surgical planning and computer assisted interventions, manually annotating data with pixel-level annotations for each test domain is not a feasible solution. Adaptation algorithms to improve the model generalizability from source domain to target domain has significant practical value. In this work, we investigate unsupervised domain adaptation (UDA) technique to train a cross-domain segmentation method which is robust to domain shift, and which does not require any annotations on the test domain.

Given respectively two sets of unpaired data in two different domains, majority of existing unsupervised cross-domain segmentation methods are based on either feature alignment (Dou et al., 2019; Wu and Zhuang, 2020) or image translation (Chartsias et al., 2017; Chen et al., 2019b; 2020b; 2021). Typically, image translation-based methods consist of two components: a cross-domain image synthesis component where adversarial learning with cycle-consistent reconstruction is leveraged to solve the prob-

* Corresponding author.
E-mail address: guoyan.zheng@sjtu.edu.cn (G. Zheng).
[1] Authors contribute equally to the paper.

https://doi.org/10.1016/j.media.2021.102328
1361-8415/

lem of data without correspondence, and an image segmentation component. The identified limitations of existing methods include: (1) most of them do not preserve semantic information in the process of image translation such that any error generated in the cross-domain image synthesis stage will be passed to the image segmentation stage; and (2) the cross-domain image synthesis stage cannot benefit from the high-level semantic information obtained from the image segmentation stage.

In this paper, we propose **Cy**cle-consistent **C**ross-domain **M**edical **I**mage **S**egmentation, referred as CyCMIS, integrating diverse image translation via disentangled representation learning and semantic consistency regularization into one network. Instead of learning one-to-one mapping as in CycleGAN (Zhu et al., 2017), our method characterizes the complex relationship between domains as many-to-many mapping (Yang et al., 2019b), where images are embedded onto two spaces: a domain-invariant content space and a domain-specific attribute space. By enforcing diversity seeking regularization (Yang et al., 2019a; Mao et al., 2019), disentangled content features can be combined with different attribute vectors to produce diverse images with identical content, which can be regarded as a way of online diverse image augmentation. Based on this, we further propose a novel diverse inter-domain semantic consistency loss to regularize the cross-domain segmentation. We additionally introduce an intra-domain semantic consistency loss to encourage the segmentation consistency between the original input and the image after cross-cycle reconstruction. Our contributions are summarized as follows:

1. We propose an end-to-end unsupervised cross-domain image segmentation framework integrating diverse image translation with semantic image segmentation into one network such that the two parts can benefit from each other, i.e., better image translation will improve cross-domain image segmentation and conversely, better image segmentation will regularize cross-domain image translation.
2. We introduce two consistency losses, i.e., the diverse inter-domain semantic consistency loss and the intra-domain semantic consistency loss, to further regularize the cross-domain segmentation process.
3. We demonstrate on two public datasets that the proposed end-to-end network, which takes both content features and diverse appearance information into account, producing better segmentation than state-of-the-art (SOTA) cross-domain segmentation methods.

2. Related work

Unsupervised domain adaptation has been applied to a large number of medical image analysis tasks (Choudhary et al., 2020; Guan and Liu, 2021). The existing methods can be largely categorized into two groups: shallow learning-based methods and deep learning-based methods. Below we will give a review of related work.

2.1. Shallow learning-based methods

The methods in this group are built on human-engineered features and conventional machine learning models (Guan and Liu, 2021). They learn to minimize the discrepancy among data distributions in different domains. For example, Heimann et al. (2013) proposed to use instance weighting to adapt a probabilistic boosting-tree-based ultrasound transducer localization approach, which was trained on *in silico* simulation data that could be generated in great quantities with perfectly accurate labels, to *in vivo* fluoroscopy data. The instance weights were calculated with logistic regression. Another instance weighting strategy has been employed by Cheplygina et al. (2017) to address domain shift problem in multicenter classification of chronic obstructive pulmonary disease. Although success has been achieved by shallow learning-based UDA methods on image classification problems, it is even more trickier to deal with semantic segmentation problem, as not only the classification information (what) but also the localization information (where) matters. Few shallow learning-based UDA methods have been reported for cross-domain image segmentation.

2.2. Deep learning-based methods

Semantic segmentation is regarded as one of the most important tasks in medical image analysis (Litjens et al., 2017). Due to shift between domains, CNN models trained on one domain frequently fail on another (Choudhary et al., 2020; Guan and Liu, 2021). To mitigate the problem, deep unsupervised cross-domain medical image segmentation has drawn more and more attentions. Some methods are based on feature alignment while others are based on image translation. For example, Yan et al. (2019b) proposed a Domain-adversarial Neural Network (DANN)-based domain adaptation method for cross-vendor segmentation of left ventricle from cine MRI sequences. A domain discriminator is co-trained with a segmentation network to learn domain-invariant features for the task of segmentation. Unlike DANN (Ganin et al., 2016), where only the last layer of the task-specific network is adapted, Yan et al. (2019b) designed multi-output domain adaptor by adding additional output discriminator onto more layers (e.g., the penultimate layer). In contrast, Dou et al. (2019) proposed a plug-and-play adversarial domain adaptation network for cardiac MR and CT image segmentation by only adapting low-level layers while keeping higher layers fixed to reduce domain shift during training. Instead of implicit domain adaptation via adversarial learning, Wu and Zhuang (2020) proposed to explicitly align distributions by minimizing a distance of characteristic functions (referred as "CF Distance") in a common latent feature space and applied it to cross-domain cardiac image segmentation.

In comparison with feature alignment, image translation-based UDA approaches perform alignment in the image space instead of the latent feature space, leading to better interpretability through visual inspection of translated images. With the wide success of CycleGAN in unpaired image-to-image transformation (Zhu et al., 2017), many early unsupervised cross-domain medical image segmentation approaches are based on modified CycleGAN. For example, Chartsias et al. (2017) introduced a two-stage framework to segment cardiac MR images using annotations of CT images. Jiang et al. (2018) combined a tumor-aware unsupervised cross-domain adaptation with semi-supervised tumor segmentation using U-net (Ronneberger et al., 2015) trained with both synthesized and limited number of original MR images. Image translation and image segmentation can also be trained end-to-end (Huo et al., 2018; Liu, 2019; Chen et al., 2020b), which reported to achieve superior performance than the two-stage methods. Moreover, to address the intrinsic ambiguity of cycle-consistent reconstruction with respect to geometric transformation, previous work introduced shape consistency (Cai et al., 2019), anatomy-regularization (Chen et al., 2020b), and structural-similarity constraints (Hiasa et al., 2018).

Disentangled representation learning has also been widely researched for medical image segmentation (Yang et al., 2019b; Chartsias et al., 2019; 2020; Chen et al., 2019a; 2019b; 2021). Disentangled representation learning-based approaches (Huang et al., 2018; Lee et al., 2020) can learn to embed images onto two spaces: (1) a domain-invariant content space and (2) a domain-specific attribute space. By assuming that a shared domain-invariant content space can be found for both domains that preserves the structural information, Yang et al. (2019b) proposed to train a seg-

mentation model on content-only images from annotated CT images and then apply it directly on content-only images from unlabeled MR images. Instead of separating the image translation stage from the segmentation stage as in (Yang et al., 2019b), Chen et al. (2021) proposed an end-to-end diverse data augmentation generative adversarial network (DDA-GAN) for learning image segmentation with cross-domain annotations. Although diverse data augmentation was used to generate synthetic images in the unlabeled target domain given a single image in the annotated source domain, the segmentation model in (Chen et al., 2021) was trained on content-only images in the target domain. In contrast, instead of working on disentangled content-only images, Chen et al. (2019b) proposed an unsupervised multi-modal style transfer method to transfer anatomical knowledge and features learned on annotated balanced steady-state free precession (bSSFP) images onto segmenting cardiac structures from the late-gadolinium enhanced (LGE) cardiac images. Their method was based on multiple networks, i.e., a multi-modal image translation network (Huang et al., 2018) for attribute transfer and two cascaded segmentation networks for image segmentation. They argued for that the translation network could generate realistic and diverse synthetic LGE images given a single bSSFP image, enabling generative model-based data augmentation for improving the generalizability of the segmentation network.

There are a few methods investigating the combination of feature alignment with image alignment for an improved performance (Yan et al., 2019a; Chen et al., 2020a). Chen et al. (2020a) proposed Synergistic Image and Feature Alignment (SIFA) framework for cross-domain cardiac image segmentation. Specifically, labeled source images were first transformed into target-like image using CycleGAN. Then, adversarial learning-based feature alignment was further used to reduce the domain gap. Yan et al. (2019a) proposed a similar framework for cross-vendor cardiac cine MR image segmentation, combining feature alignment with image alignment.

3. Method

Let $x \in X$ and $y \in Y$ be images from two domains, and $m_x \in M_X$ and $m_y \in M_Y$ be corresponding labels to x and y, respectively. Note that x and y are not necessarily paired, and we have no access to M_Y in the training phase. Our goal is to design a network to segment unlabeled images in the target domain Y by making use of X and M_X in the source domain. Fig. 1 shows an overview of the proposed CyCMIS framework which consists of two modules: a diverse image translation (DIT) module and a domain-specific segmentation (DSS) module, as detailed below.

3.1. Diverse image translation

Similar to (Huang et al., 2018; Lee et al., 2020), the DIT module consists of two main components: Variational AutoEncoder (VAE) for reconstruction and Generative Adversarial Networks (GAN) for adversarial learning. The VAE components are trained for intra-domain reconstruction, where the reconstruction loss is minimized to encourage the encoders and generators to invert one another. The GAN components are trained for two purposes. The first one is for disentangled representation learning, i.e., the input space is decomposed into domain-invariant content and domain-specific attribute subspaces. The second one is to promote diverse cross-domain synthesized images as realistic as possible. As shown in Fig. 1, the DIT module consists of two content encoders $\{E_x^c, E_y^c\}$, two attribute encoders $\{E_x^a, E_y^a\}$, two generators $\{G_x, G_y\}$, a content feature discriminator D^c, and two domain discriminators $\{D_x, D_y\}$. For given images x and y, we can obtain the disentangled content features $c_x = E_x^c(x)$ and $c_y = E_y^c(y)$, respectively, and the corresponding attribute codes $a_x = E_x^a(x)$ and $a_y = E_y^a(y)$, which are empirically set as a 8-bit vector. The attribute encoder is trained to embed images into a latent space $a_d, d = \{x, y\}$ that matches the estimated distribution of $p(a_d)$ to the Gaussian distribution $N(0, \mathbf{I})$ by minimizing Kullback-Leibler (KL) divergence of $L_{KL} = \mathbb{E}_{a_d}[D_{KL}(p(a_d)||N(0, \mathbf{I}))]$. Then the translated image x' in the source domain is generated by combining the content feature from y and the attribute code from x, i.e., $x' = G_x(c_y, a_x)$. Similarly, we can generate $y'_1 = G_y(c_x, a_y)$ in the target domain by combining the content feature from x and the attribute code from y. The above process can be used further to obtain the cross-cycle reconstruction image $\hat{x} = G_x(c'_y = E_y^c(y'_1), a'_x = E_x^a(x'))$ and $\hat{y} = G_y(c'_x = E_x^c(x'), a'_y = E_y^a(y'_1))$, which are expected to have the same content and attribute as x and y, respectively. Note that to achieve diverse image translation in the target domain, the content feature c_x is combined with multiple attributes codes sampled from the Gaussian distribution $N(0, \mathbf{I})$. The diversely generated images can be represented as $\{y'_i = G_y(c_x, z_i)\}$, where $z_i \sim N(0, \mathbf{I}), i \in \{2, 3, \ldots, n\}$. In addition, a diversity-seeking regularization is incorporated to encourage the generator G_y to be diversity-sensitive.

3.1.1. Disentangled representation learning via self and cross-cycle reconstruction

As shown in Fig. 1, the disentangled content feature and attribute code are combined for image-to-image translation. We leverage the property that a translated image in a domain to be similar to its original version, which has two different situations:

- Cross-cycle reconstruction, when the content features and the attribute features are from images in different domains, and
- Self-reconstruction, when the content features and the attribute features are from the same image.

Accordingly, we introduce intra-domain reconstruction loss which is composed of self-reconstruction loss L_{self_rec}, and cross-cycle reconstruction loss L_{cc_rec} in both source and target domains, i.e., $L_{recon} = L^x_{self\_rec} + L^y_{self\_rec} + L^x_{cc\_rec} + L^y_{cc\_rec}$, where each item is defined as follows:

$$L^x_{self\_rec}(E_x^c, E_x^a, G_x) = \|G_x(E_x^c(x), E_x^a(x)) - x\|_1 \tag{1}$$

$$L^y_{self\_rec}(E_y^c, E_y^a, G_y) = \left\|G_y(E_y^c(y), E_y^a(y) - y\right\|_1 \tag{2}$$

$$L^x_{cc\_rec}(E_x^c, E_x^a, E_y^c, E_y^a, G_x, G_y) = \left\|\hat{x} - x\right\|_1 \tag{3}$$

$$L^y_{cc\_rec}(E_x^c, E_x^a, E_y^c, E_y^a, G_x, G_y) = \left\|\hat{y} - y\right\|_1 \tag{4}$$

Where $\hat{x} = G_x(E_y^c(G_y(E_x^c(x), E_y^a(y))), E_x^a(G_x(E_y^c(y), E_x^a(x))))$, and $\hat{y} = G_y(E_x^c(G_x(E_y^c(y), E_x^a(x))), E_y^a(G_y(E_x^c(x), E_y^a(y))))$.

Additionally, in order to encourage invertible mapping between the image and the latent space, the disentangled content feature and attribute code of a translated image should be similar to the original content feature and attribute code that are used to generate the image. To this end, we introduce latent space reconstruction loss which consists of content space reconstruction loss and attribute space reconstruction loss in both source and target domains, i.e., $L_{lat\_rec} = L^x_{lat\_cont\_rec} + L^y_{lat\_cont\_rec} + L^x_{lat\_app\_rec} + L^y_{lat\_app\_rec}$, where each item is defined as follows:

$$L^x_{lat\_app\_rec}(E_y^c, E_x^a, G_x) = ||a'_x - E_x^a(x)||_1 \tag{5}$$

$$L^y_{lat\_app\_rec}(E_x^c, E_y^a, G_y) = ||a'_y - E_y^a(y)||_1 \tag{6}$$

$$L^x_{lat\_cont\_rec}(E_x^c, E_y^c, E_x^a, G_y) = ||c'_y - E_x^c(x)||_1 \tag{7}$$

$$L^y_{lat\_cont\_rec}(E_x^c, E_y^c, E_y^a, G_x) = ||c'_x - E_y^c(y)||_1 \tag{8}$$

Were $a'_x = E_x^a(G_x(E_y^c(y), E_x^a(x)))$, $a'_y = E_y^a(G_y(E_x^c(x), E_y^a(y)))$, $c'_y = E_y^c(G_y(E_x^c(x), E_y^a(y)))$, and $c'_x = E_x^c(G_x(E_y^c(y), E_x^a(x)))$.

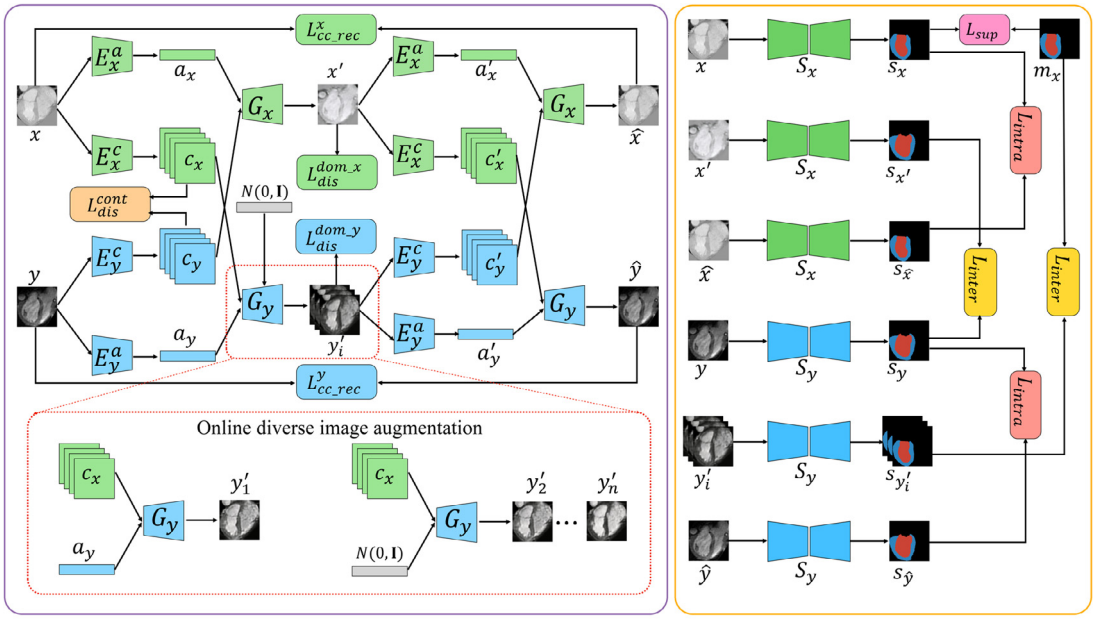

Fig. 1. A schematic illustration of our proposed CyCMIS framework which consists of (a) diverse image translation module; and (b) domain specific segmentation module. Green and blue colored blocks show parts related with source domain and target domain, respectively. The dotted red rectangle shows the online diverse image augmentation block. See main text for detailed explanation of our framework. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.1.2. Generative adversarial learning for domain-invariant content space

We encourage the disentangled content features $c_x = E_x^c(x)$ and $c_y = E_y^c(y)$ to be domain-invariant, which is realized by a generative adversarial learning strategy with a content feature discriminator D^c to differentiate between content features of different domains. The encoders E_x^c, E_y^c and the content feature discriminator D^c are trained alternatively by playing a minimax game. Specifically, the content feature discriminator D^c is trained to learn to judge whether content features are of the source domain or of the target domain by minimizing following loss function:

$$L_{dis}^{cont}(D^c) = -\mathbb{E}_x[\log(1 - D^c(E_x^c(x)))] - \mathbb{E}_y[\log(D^c(E_y^c(y)))] \quad (9)$$

During the encoder training stage, E_x^c and E_y^c are trained to extract content features that are indistinguishable from each other by minimizing following loss function:

$$L_{adv}^{cont}(E_x^c, E_y^c) = -\mathbb{E}_x\left[\frac{1}{2}\log D^c(E_x^c(x)) + \frac{1}{2}\log(1 - D^c(E_x^c(x)))\right] - \mathbb{E}_y\left[\frac{1}{2}\log D^c(E_y^c(y)) + \frac{1}{2}\log(1 - D^c(E_y^c(y)))\right] \quad (10)$$

3.1.3. Generative adversarial learning for diverse image translation

Following (Mao et al., 2019), we enforce diversity seeking regularization to alleviate the mode-collapse problem during online data augmentation in the target domain with loss:

$$L_{div}(E_x^c, G_y) = -\frac{\|G_y(E_x^c(x), z_2) - G_y(E_x^c(x), z_3)\|_1}{\|z_2 - z_3\|_1} \quad (11)$$

Where $(z_2, z_3) \sim N(0, \mathbf{I})$ as described above.

Additionally, to encourage the diverse cross-domain synthesized images to be as realistic as possible, two domain discriminators D_x and D_y are trained to distinguish between the real and synthesized images in the source domain and the target domain, respectively, by minimizing following loss function:

$$L_{dis}^{dom}(D_x, D_y) = L_{dis}^{dom\_x}(D_x) + \sum_{i=1}^{n} L_{dis}^{dom\_y_i}(D_y) \quad (12)$$

where:

$$L_{dis}^{dom\_x}(D_x) = -\mathbb{E}_x[logD_x(x)] - \mathbb{E}_{x'=G_x(c_y,a_x)}[log(1 - D_x(x'))] \quad (13)$$

$$L_{dis}^{dom\_y_i}(D_y) = -\mathbb{E}_y[logD_y(y)] - \mathbb{E}_{y_i'=G_y(c_x,z_i)}[log(1 - D_y(y_i'))] \quad (14)$$

The generators G_x and G_y are trained to synthesize realistic images to fool the discriminators by minimizing loss function:

$$L_{adv}^{dom}(G_x, G_y) = -L_{dis}^{dom}(D_x, D_y) \quad (15)$$

The overall objective loss function of the DIT module is:

$$L_{D_x,D_y,D^c} = L_{dis}^{dom} + L_{dis}^{cont} \quad (16)$$

$$L_{E_x^c,E_y^c,E_x^a,E_y^a,G_x,G_y} = (L_{adv}^{dom} + L_{adv}^{cont}) + \lambda_1 L_{recon} + \lambda_2 L_{lat\_rec} + \lambda_3 L_{KL} \quad (17)$$

where λ_1, λ_2 and λ_3 are parameters controlling the relatively weights of different losses, which will be detailed below.

3.2. Domain-specific segmentation

The DSS module contains two segmentation networks, S_x and S_y, for the source domain and the target domain, respectively. The main idea of DSS is to use supervised source domain segmentation to guide unsupervised target domain segmentation, via the diverse image translation. To this end, we proposed a new consistency loss L_{con} to regularize the cross-domain segmentation process in addition to the supervised learning in the source domain. The overall loss function for the DSS module is:

$$L_{DSS} = L_{sup} + \mu L_{con} \quad (18)$$

where μ is a parameter controlling the relative weights between supervised learning and consistency regularization.

Before we describe details on L_{sup} and L_{con}, we first present the segmentation loss function L_{seg} used below, which combines cross entropy loss with Dice loss and was originally introduced in (Patravali et al., 2017):

$$L_{seg}(\alpha, \beta, S) = -\sum_{k=1}^{K}\left[\beta_j^k \log(S(\alpha)_j^k) + \frac{2\sum_{j=1}^{N}\beta_j^k \hat{S}(\alpha)_j^k}{\sum_{j=1}^{N}\beta_j^k\beta_j^k + \hat{S}(\alpha)_j^k\hat{S}(\alpha)_j^k}\right] \quad (19)$$

where α denotes the input image, β refers to the reference segmentation (please note, the reference segmentation could be the reference annotations, or predicted segmentation), S represents the segmentation network. $S(\alpha)_j^k$ and $\hat{S}(\alpha)_j^k$ respectively denote the probability prediction and the one-hot output of voxel j for class k.

3.2.1. Supervised segmentation loss

The supervised segmentation loss is defined on the source domain as:

$$L_{sup} = \mathbb{E}_x[L_{seg}(x, m_x, S_x)] \tag{20}$$

3.2.2. Semantic consistency loss

The semantic consistency loss consists of a novel diverse inter-domain semantic consistency loss L_{inter} and an intra-domain semantic consistency loss L_{intra}. Specifically, the original image x and the diversely translated image $\{y_i'\}(i = 1, 2, \ldots, n)$ should have the same semantic structures although they belong to different domains. The diverse image generation plays a role of on-line image augmentation, aiming to improve the accuracy and robustness of cross-domain segmentation. L_{inter} is used to constrain the segmentation results between x and $\{y_i'\}(i = 1, 2, \ldots, n)$ to be consistent, and between y and x' to be consistent. Similarly, L_{intra} is used to enforce the segmentation consistency between the original input (x or y) and the cross-cycle reconstructed image ($\hat{x}$ or $\hat{y}$). The consistency loss is represented as $L_{con} = L_{inter} + L_{intra}$, where each item is defined as follows:

$$L_{inter}(E_x^g, E_y^a, G_y, S_y) = \left(\sum_{i=1}^{n} \mathbb{E}_x[L_{seg}(y_i', m_x, S_y)]\right) + \mathbb{E}_y[L_{seg}(y, S_x(x'), S_y)] \tag{21}$$

$$L_{intra}(E_x^g, E_x^a, E_y^g, E_y^a, G_x, G_y, S_x, S_y) = \mathbb{E}_x[L_{seg}(\hat{x}, S_x(x), S_x)] + \mathbb{E}_y[L_{seg}(\hat{y}, S_y(y), S_y)] \tag{22}$$

3.3. Training strategy

The training strategy is detailed in Algorithm 1. At each iteration, we first train content discriminator D^c for $(D_{step}^c - 1)$ steps. After that, we fix content discriminator D^c and train other components step by step as shown in Algorithm 1. At each step, we conduct following computations including forward propagation, loss calculation, backward propagation and updating weights of associated components. Above procedure is repeated until the total number of iterations reaches T=100,000. We empirically set $D_{step}^c = 3$ such that content discriminator D^c gets trained more often than other components in order to facilitate domain-invariant content alignment.

Algorithm 1 CyCMIS Training strategy.

GET Image set X and Y, and label set M_x corresponding to X
FOR t=1 to T
  IF t % D_{step}^c != 0
    Train D^c with loss L_{dis}^{cont}
    CONTINUE
  END IF
  Train D_x, D_y with loss L_{dis}^{dom}
  Train E_x^c, E_y^c with loss L_{adv}^{cont}
  Train G_x, G_y with loss L_{adv}^{dom}
  Train $E_x^c, E_x^a, E_y^c, E_y^a, G_x, G_y$ with loss $\lambda_1 L_{recon} + \lambda_2 L_{latent} + \lambda_3 L_{KL}$
  Train S_x with loss L_{sup}
  Train E_x^c, E_y^a, G_y, S_y with loss $\mu \sum_{i=1}^{n} L_{inter}^i$
  Train $E_x^c, E_x^a, E_y^c, E_y^a, G_x, G_y, S_x, S_y$ with loss μL_{intra}
END FOR

3.4. Implementation details

The proposed CyCMIS framework was implemented in PyTorch and trained with a Tesla V100 graphics card. As shown in Fig. 2, the content encoders consist of three convolutional layers and four residual blocks followed by batch normalization (He et al., 2016), while attribute encoders contain five convolutional layers followed by a fully connected layer. The generators are composed of two convolutional layers followed by three transposed convolutional layers The discriminators are composed of convolutional layers for binary classification. The segmentation networks are a standard PSPNet (Zhao et al., 2017). For training, we used the Adam optimizer with a batch size of 2, and exponential decay rates $(\beta_1, \beta_2) = (0.5, 0.999)$. The initial learning rate was set to 0.0001 and 0.001 for DIT and DSS, respectively. We empirically set $\lambda_1 = 10$, $\lambda_2 = 10$, $\lambda_3 = 0.01$ and $\mu = 1$. Training of CyCMIS model took on average 367 minutes on a Tesla V100 graphics card.

4. Experiments and results

In this section, we present the experimental results of the proposed CyCMIS framework. Two publicly available image datasets were used in our study: cardiac MR dataset from the Multi-Sequence Cardiac MR Segmentation (MS-CMRSeg) challenge (Zhuang, 2018) and CT-MR dataset from the Multi-Modality Whole Heart Segmentation (MM-WHS) challenge (Zhuang et al., 2019). The MS-CMRSeg challenge dataset contains 45 paired bSSFP CMR and LGE CMR images with ground truth annotations while the MM-WHS challenge dataset contains 20 unpaired CT and MR images with ground truth annotations. The reason why we choose these two datasets is because they differ in nature and may post different levels of challenge to UDA. Specifically, on the MM-WHS challenge dataset, we have to deal with large domain shift caused by different modalities (i.e., cross-modality), whose underlying imaging physics is completely different, though the background of the dataset is relatively less complex. In contrast, domain shift of the MS-CMRSeg challenge dataset is caused by different MR imaging protocols (i.e., cross-protocol) and thus is relatively smaller. However, the background of the dataset is relatively more complex than that of the other dataset.

4.1. Task and evaluation metrics

For the MM-WHS challenge dataset, our goal is to train a network in target domain for segmenting the left ventricular cavity (LV) and the left ventricular myocardium (MYO) by transferring anatomical knowledge and features learned on annotated source domain. Following the practice introduced in (Wu and Zhuang, 2020), we sampled 16 slices from each 3D image along the long-axis view around the center of LV. Then these slices were cropped around the heart and scaled to a size of 224×224 pixels. For the MS-CMR challenge dataset, considering the fact that while the borders of the myocardium is difficult to delineate on LGE CMR images, they are clear and easy to identify on the bSSFP CMR images (Chen et al., 2021). We are aiming to segment the LV, the MYO and the right ventricle blood cavity (RV) from the LGE CMR images (the target domain) by transferring knowledge and features learned on annotated bSSFP images (the source domain). Each bSSFP CMR image contains 8 to 12 contiguous slices with in-plane resolution of $1.25mm \times 1.25mm$, while each LGE CMR images consists of 10 to 18 slices with in-plane resolution of $0.75mm \times 0.75mm$. The slices of each modality cover the main body of the ventricles. We shuffled all slices to make the training data unpaired and then scaled them to 224×224 pixels.

On each challenge dataset, we conducted a standard 5-fold cross validation study. Specifically, data of target domain of each

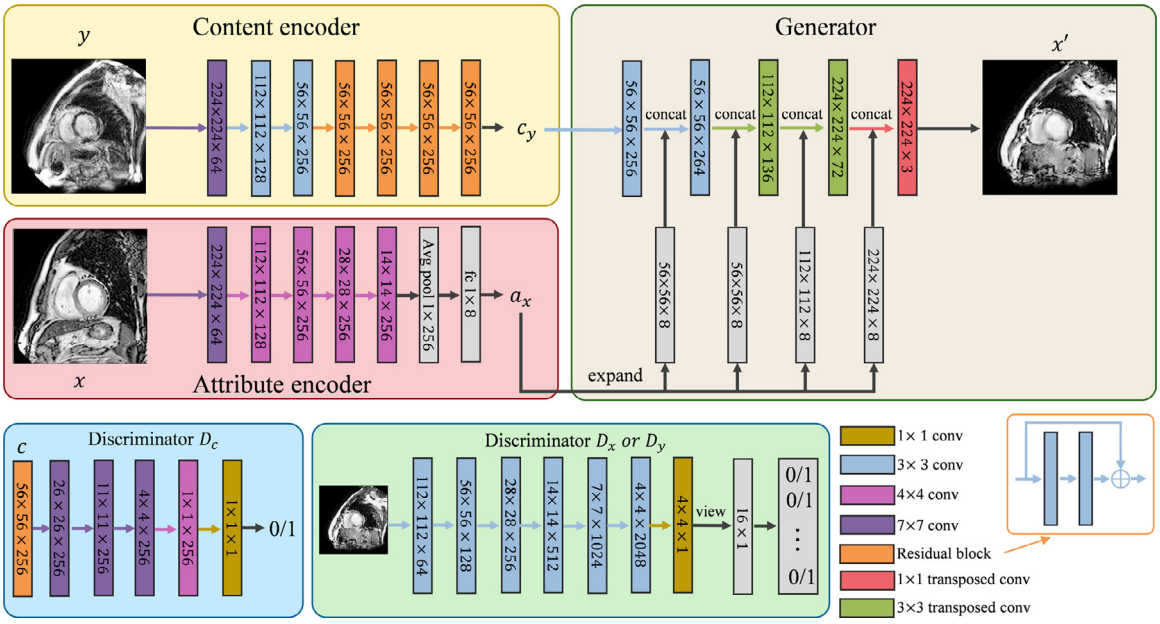

Fig. 2. Network architectures for encoders, generator and discriminators used in our framework.

dataset were randomly partitioned into 5 equal size subsamples depending on subjects. Each time, a single subsample was used as testing data while the remaining 4 subsamples were used as training data. This process was repeated 5 times, with each one of the 5 subsamples used exactly once as the test data. In each fold, segmentation performance was quantified using the Dice Similarity Coefficient (DSC) and the Average Symmetric Surface Distance (ASSD) in 3D after transforming predicted results to the original space. When there is no prediction for a class, the ASSD for that class is taken as the diagonal length of the associated data volume. Both DSC and ASSD have been used in previous work to evaluate the performance of different unsupervised cross-domain medical image segmentation methods (Chen et al., 2020b; 2021; Wu and Zhuang, 2020).

We compared our method with state-of-the-art UDA approaches including CycleGAN (Zhu et al., 2017), DRIT++ (Lee et al., 2020), AdaptSegNet (Tsai et al., 2018), DDA-GAN (Chen et al., 2021), and SIFA (Chen et al., 2020a), which leveraged disentangled representation learning, feature alignment, image alignment or a combination of both alignment for unsupervised domain adaptation. In addition, we further compared our method with two other state-of-the-art UDA methods incorporating semantic consistency regularization: CyCADA (Hoffman et al., 2018) and ICMSC (Zeng et al., 2020). CycleGAN, DRIT++, AdaptSegNet and CyCADA are all well-established UDA method on natural image datasets while DDA-GAN, SIFA and ICMSC have been applied to cross-modality segmentation on MR/CT images. All competing methods were trained and evaluated using the same training and testing dataset as our proposed approach. For CycleGAN[2], DRIT++[3], AdaptSegNet[4], SIFA[5], and CyCADA[6], we adopted the official implementations as provided by the associated authors. We implemented the method described in (Zeng et al., 2020) and in (Chen et al., 2021). For CyCADA, AdaptSegNet, DDA-GAN, and SIFA, where segmentation is part of the methods, we used the segmentation network of the associated method. For others, for a fair comparison, the segmentation network is the same as that used in our proposed approach.

4.2. Experimental results

4.2.1. Results on the MM-WHS challenge dataset

Effectiveness of CyCMIS on cross-modality image segmentation. We first investigated the influence of domain shift on cross-modality image segmentation. The influence was demonstrated by comparing the results of the same segmentation network (Zhao et al., 2017) when evaluated on the same testing images of the target domain but trained with different data: (1) the first model was trained on images and the associated annotations of the source domain; and (2) the second model was trained on training images and the associated annotation of the target domain. The first model could be regarded as the "No adaptation" lower bound without using any domain adaptation technique while the second model was the "Full supervision" upper bound when the segmentation network was trained using annotations of the target domain. The performance gap between these two models can be used to measure the influence of domain shift on cross-modality image segmentation. The results are presented in Table 1. The "No adaptation" model trained on CT images only obtained an average DSC of 16.70% and an average ASSD of 16.64mm when being tested on MR images directly. In contrast, the "Full supervision" model achieved an average DSC of 87.25% and an average ASSD of 0.82mm, demonstrating the significant influence of domain shift on cross-modality image segmentation. When evaluated along the reverse direction, a similar influence was also observed. Specifically, the "No adaptation" model obtained an average DSC of 32.80% and an average ASSD of 35.32mm while the "Full supervision" model achieved an average DSC of 83.99% and an average ASSD of 1.25mm.

We further investigated the effectiveness of CyCMIS on cross-modaity image segmentation. As shown in Table 1, when taking CT images as the source domain and MR images as the target domain, CyCMIS achieved an average DSC of 84.51% and an average ASSD of 1.00mm, where were close to the results achieved by the "Full supervision" upper bound model, demonstrating the effectiveness of CyCMIS. When evaluated along the reverse direction, a

[2] https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix.
[3] https://github.com/HsinYingLee/DRIT.
[4] https://github.com/wasidennis/AdaptSegNet.
[5] https://github.com/cchen-cc/SIFA.
[6] https://github.com/jhoffman/cycada_release.

Table 1
Quantitative comparison between our method and the other SOTA methods on the MM-WHS challenge dataset.

Cardiac CT → Cardiac MR						
Methods	DSC(%)			ASSD(mm)		
	LV	MYO	Average	LV	MYO	Average
No adaptation	27.77	5.6	16.70	24.27	9.01	16.64
Full supervision	92.09	82.41	87.25	0.66	0.97	0.82
CycleGAN (Zhu et al., 2017)	63.18	57.32	60.25	4.28	3.38	3.83
DRIT++ (Lee et al., 2020)	75.09	58.31	66.70	3.31	2.86	3.09
AdaptSegNet (Tsai et al., 2018)	80.46	56.74	68.60	1.64	2.28	1.96
CyCADA (Hoffman et al., 2018)	72.25	68.53	70.39	3.56	2.86	3.21
ICMSC (Zeng et al., 2020)	75.91	73.10	74.51	3.28	2.23	2.75
DDA-GAN (Chen et al., 2021)	86.28	74.74	80.51	1.34	1.35	1.34
SIFA (Chen et al., 2020a)	86.43	75.86	81.14	1.54	1.48	1.51
CyCMIS (ours)	**90.11**	**78.91**	**84.51**	**0.91**	**1.08**	**1.00**
Cardiac MR → Cardiac CT						
Methods	DSC(%)			ASSD(mm)		
	LV	MYO	Average	LV	MYO	Average
No adaptation	47.48	18.12	32.80	41.83	44.67	35.32
Full supervision	86.94	81.04	83.99	1.40	1.10	1.25
CycleGAN (Zhu et al., 2017)	75.92	52.36	64.14	2.25	3.49	2.87
AdaptSegNet (Tsai et al., 2018)	76.74	55.79	66.30	3.27	2.80	3.03
DRIT++ (Lee et al., 2020)	79.42	56.96	68.19	2.92	3.19	3.06
CyCADA (Hoffman et al., 2018)	78.31	59.14	68.72	1.93	2.79	2.36
ICMSC (Zeng et al., 2020)	83.07	68.56	75.81	1.17	1.73	1.45
DDA-GAN (Chen et al., 2021)	85.88	70.85	78.36	1.66	2.16	1.91
SIFA (Chen et al., 2020a)	88.02	74.06	81.04	1.06	1.57	1.32
CyCMIS (ours)	**90.11**	**79.44**	**84.77**	**0.88**	**1.09**	**0.98**

similar performance improvement was observed. Specifically, CyCMIS achieved an average DSC of 84.77% and an average ASSD of 0.98mm, which were even better than the results achieved by the "Full supervision" upper bound model, especially on LV. The reasons could be as follows. The "Full supervision" upper bound model each time was trained with only the labeled training images in the target domain. By contrast, as shown in Eq. (21) and (22), the segmentation network S_y of the present approach in the target domain was trained with the diversely translated images y_i' with labels m_x which were generated by combining the content features of the source domain with multiple attribute codes of the target domain, the target domain images y with pseudo-labels $S_x(x')$, and the cross-cycle reconstructed images $\hat{y}$ with pseudo-labels $S_y(y)$. Thus, the training set is larger and more diverse. The reasons why such an enlarged and diversely training dataset is possible are because (a) the present approach uses the disentangled representation framework for learning to generate diverse outputs with unpaired data. As shown in Fig. 3, the shared content space provided a robust representation for structural information in different domains; and (b) we introduce two semantic consistency losses to regularize cross-modality segmentation.

The qualitative segmentation results in Fig. 4 also show that without adaptation, it is difficult to identify any valid cardiac structures from the segmentation results. In contrast, CyCMIS can segment clinically meaningful cardiac structures even when there are ambiguous boundaries between the target structures and the surrounding tissues. Both qualitative and quantitative results demonstrate the effectiveness of CyCMIS on overcoming the severe domain shift when applied to cross-modality adaptation.

Comparison with the SOTA methods. Table 1 also shows the comparison of CyCMIS with the seven SOTA methods. Along both directions, DRIT++ (Lee et al., 2020), achieved better results than CycleGAN (Zhu et al., 2017), indicating the effectiveness of disentangled representation learning in addressing domain shift. DDA-GAN (Chen et al., 2021) achieved much better results than DRIT++ (Lee et al., 2020) demonstrating the effec-

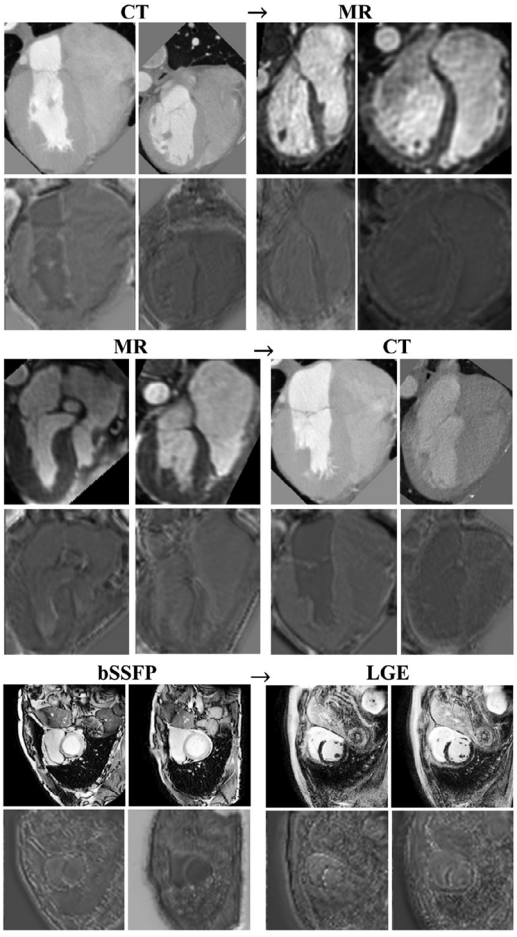

Fig. 3. Examples of content-only images via disentangled representation for different unsupervised cross-domain segmentation scenarios; Top two rows: CT to MR; middle two rows: MR to CT; and bottom two rows: bSSFP to LGE. For each scenario, the first row shows the original images while the second row presents the content-only images.

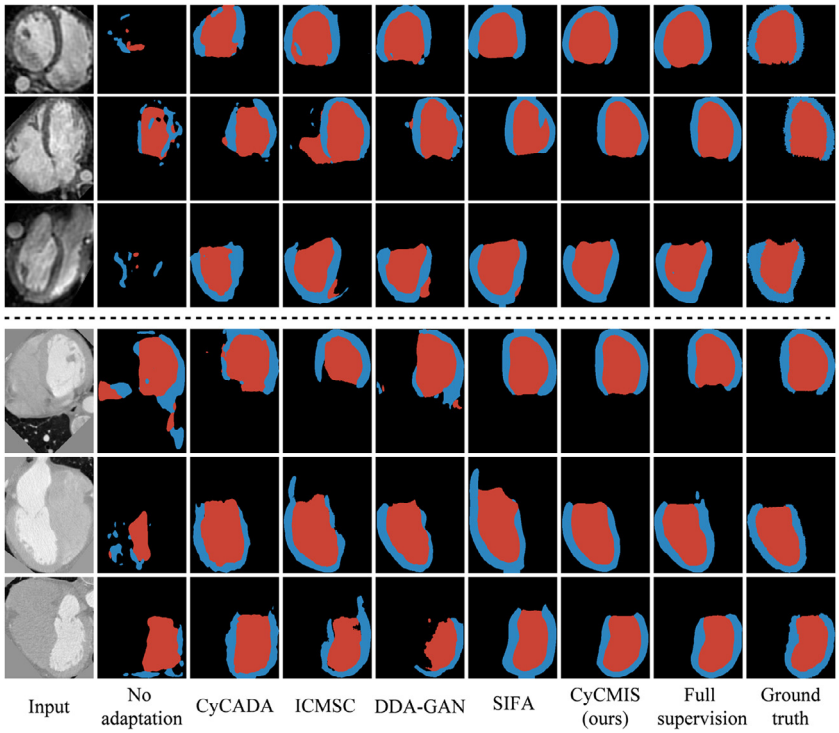

Fig. 4. Qualitative comparison of segmentation results by different methods on MM-WHS challenge dataset; top three rows: when MR imaging was chosen as the target domain; bottom three rows: when CT imaging was chosen as the target domain.

tiveness of integrating disentangled representation learning and semantic segmentation into one end-to-end network. Both CyCADA (Hoffman et al., 2018) and ICMSC (Zeng et al., 2020) achieved better results than CycleGAN (Zhu et al., 2017), indicating the effectiveness of cycle-consistent segmentation regularization on improving cross-modality image segmentation. AdaptSegNet (Tsai et al., 2018) leveraging multi-level adversarial learning for domain adaptation, achieved an average DSC of 68.60% and 66.30% when evaluated on MR images and CT images, respectively. In comparison, by synergistically using both image and feature alignment, SIFA (Chen et al., 2020a) achieved much better results along both directions than other six methods. Overall, CyCMIS achieved the best results in terms of both average DSC and ASSD. In comparison with SIFA, the average DSC achieved by CyCMIS increased 3.37% and 3.73%, and the average ASSD decreased 0.51mm and 0.34mm, when evaluated on MR images and CT images, respectively, indicating the effectiveness of integrating disentangled representation learning with semantic consistency regularization on overcoming the severe domain shift. Fig. 4 shows a qualitative comparison of CyCMIS with the top-4 SOTA methods. Both qualitative and quantitative results demonstrate the superior performance of CyCMIS in comparison with the SOTA methods.

4.2.2. Results on the MS-CMRSeg challenge dataset

Effectiveness of CyCMIS on cross-protocol image segmentation. We first investigated the influence of the domain shift between the bSSFP CMR images and the LGE CMR images on cross-protocol image segmentation. We followed the same procedure as we did on the MM-WHS challenge dataset. We trained two models, i.e., one was the "No adaptation" lower bound model trained on the bSSFP CMR images and the associated annotations when being tested on the LGE CMR images without using any domain adaptation technique while the other was the "Full supervision" upper bound model when the segmentation network was trained using annotations of the LGE CMR images. Results of these two models are presented in Table 2. From this table, one can see that the domain shift between the bSSFP CMR images and the LGE CMR images has a significant impact on cross-protocol image segmentation. Specifically, the "No adaptation" model obtained an average DSC of 44.38% and an average ASSD of 38.63mm while the "Full supervision" model achieved an average DSC of 82.53% and an average ASSD of 1.59mm. We further investigated the effectiveness of CyCMIS on cross-protocol image segmentation. As shown in Table 2, CyCMIS achieved an average DSC of 79.08% and an average ASSD of 1.68mm. The qualitative segmentation results in Fig. 5 show that without adaptation, it is difficult to identify any clinically meaningful cardiac structures from the segmentation results. In contrast, CyCMIS can segment all three cardiac structures in a clinically meaningful way even when there are low contrast and analogous intensity distributions between the targeted structures and surrounding tissues. Both qualitative and quantitative results demonstrate the effectiveness of CyCMIS on overcoming domain shift when applied to cross-protocol image adaptation.

Comparison with the SOTA methods. We compared CyCMIS with the same seven SOTA methods as we compared on the MM-WHS challenge dataset study. Quantitative and qualitative comparison results are presented in Table 2 and Fig. 5, respectively. ICMSC and SIFA achieved much better results than other five SOTA methods, indicating the effectiveness of cycle-consistent semantic segmentation as well as the synergistic image and feature alignment on addressing the domain shift for cross-protocol image segmentation. The best results were achieved by the proposed CyCMIS method, with an increase of 4.03% average DSC and a decrease of 1.06mm average ASSD when compared with the ICMSC method, demonstrating the power of integrating disentangled representation learning with semantic consistency regularization on improving performance of cross-protocol image segmentation. Both qualitative and quantitative results demonstrate the efficacy of the proposed CyCMIS method over the SOTA methods.

Table 2
Quantitative comparison between our method and the other SOTA methods on the MS-CMR challenge dataset.

Methods	DICE(%)				ASSD(mm)			
	LV	MYO	RV	Average	LV	MYO	RV	Average
No adaptation	53.65	30.95	48.55	44.38	39.49	37.30	39.11	38.63
Full supervision	89.90	76.25	81.44	82.53	1.47	1.43	1.85	1.59
CycleGAN (Zhu et al., 2017)	72.30	48.24	65.91	62.15	6.86	4.31	5.92	5.93
AdaptSegNet (Tsai et al., 2018)	73.51	49.83	66.34	63.23	8.64	3.69	5.96	6.10
CyCADA (Hoffman et al., 2018)	74.54	52.47	66.80	64.07	4.75	4.38	6.85	5.33
DDA-GAN (Chen et al., 2021)	76.95	46.37	70.37	64.56	3.02	5.91	5.22	4.72
DRIT++ (Lee et al., 2020)	78.16	53.67	66.39	66.07	4.45	3.35	3.50	3.77
SIFA (Chen et al., 2020a)	84.21	65.61	74.82	74.88	1.99	2.15	3.24	2.46
ICMSC (Zeng et al., 2020)	83.90	65.52	75.74	75.05	2.76	2.30	3.16	2.74
CyCMIS (ours)	**87.15**	**71.38**	**78.72**	**79.08**	**1.28**	**1.46**	**2.30**	**1.68**

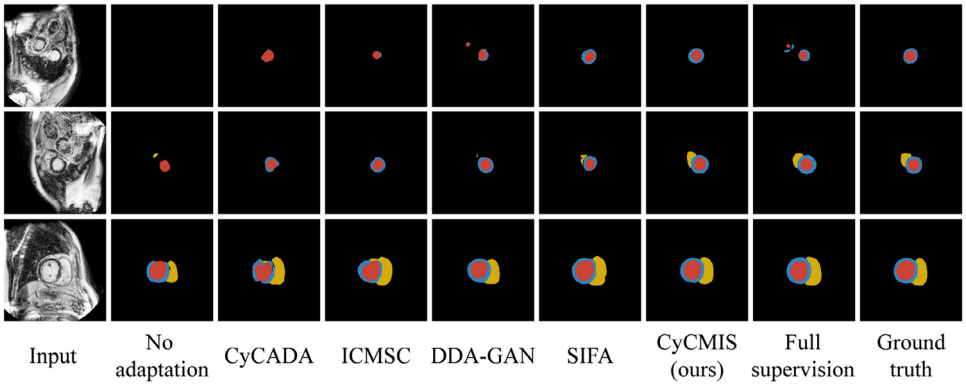

Fig. 5. Qualitative comparison of segmentation results by different methods on the MS-CMR challenge dataset.

4.2.3. Ablation study results

Influence of the number n in diverse image augmentation on image segmentation. For the proposed online diverse image augmentation, it is important to determine an optimal value of the number of images n. On one side, smaller n may limit the diversity of the image augmentation. On the other side, larger n may cause the translated images too diverse which may negatively affect the later segmentation. Additionally, larger n also means longer computation time and larger GPU memory consumption. To investigate the influence of the number n in diverse image augmentation on image segmentation, we conducted a study on both the MM-WHS challenge dataset and the MS-CMRSeg challenge dataset. For the MM-WHS challenge dataset, we took the CT images as the source domain and the MR images as the target domain. We used 4 MR images as testing images and the remaining 16 MR images as training data from the target domain. For the MS-CMRSeg challenge dataset, we took the bSSFP CMR images as the source domain and the LGE CMR images as the target domain. We used 9 LGE MR images as testing images and the remaining 36 LGE MR images as training data from the target domain. We ranged n from 1 to 5 and evaluated CyCMIS on both datasets taking DSC as the metric. The quantitative results of the study are presented in Table 3. It can be seen that when $n = 3$, CyCMIS achieved the best results on both datasets. Based on the investigation, we fixed $n = 3$ for all the experiments that we reported above. Fig. 6 shows examples of diverse image augmentation where a disentangled content structure is combined with three different attribute latent vectors to produce diverse images with identical content.

Influence of different components of CyCMIS on image segmentation. Based on the same two datasets as used in the last study, we further investigated the influence of different components of CyCMIS on image segmentation. We used DSC as the evaluation metric. The quantitative results of this study are presented

Table 3
Quantitative results of the study investigating the influence of the number n in diverse image augmentation on the cross-domain segmentation.

Methods	MM-WHS challenge dataset			MS-CMR challenge dataset			
	LV	MYO	Average	LV	MYO	RV	Average
n = 1	86.92	73.64	80.28	84.05	**71.83**	63.66	73.18
n = 2	89.02	72.84	80.93	85.92	69.44	68.32	74.56
n = 3	**90.82**	**76.33**	**83.58**	**87.11**	70.75	**79.24**	**79.03**
n = 4	87.03	74.23	80.63	86.54	68.33	71.54	75.47
n = 5	87.84	74.93	81.39	83.59	67.03	72.77	74.46

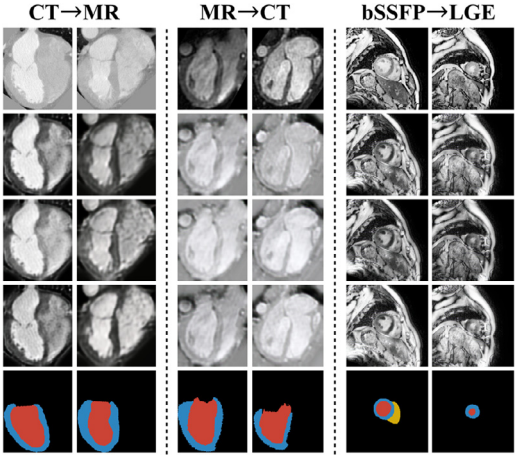

Fig. 6. Examples of diverse image augmentation; left two columns: CT to MR image augmentation examples; middle two columns: MR to CT augmentation examples; right two columns: bSSFP to LGE image augmentation examples. For each column, the top row shows the input images in the source domain while the middle three rows present the generated images in the target domain; the bottom row shows the corresponding ground truth segmentation.

Table 4
Quantitative results of the study investigating the influence of different components of the proposed CyCMIS method on the cross-domain segmentation.

Methods	L_{inter}	L_{intra}	Diversity	MM-WHS challenge dataset			MS-CMR challenge dataset			
				LV	MYO	Average	LV	MYO	RV	Average
CyCMIS	✓	✓	✓	**90.82**	**76.33**	**83.58**	**87.11**	70.75	**79.24**	**79.03**
W/o diversity	✓	✓	-	86.92	73.6	80.28	84.05	**71.83**	63.66	73.18
W/o L_{intra}	✓	-	-	84.18	70.54	77.36	83.27	65.81	60.43	69.84
W/o L_{inter}	-	-	-	79.00	57.63	68.32	68.37	45.29	52.30	55.32

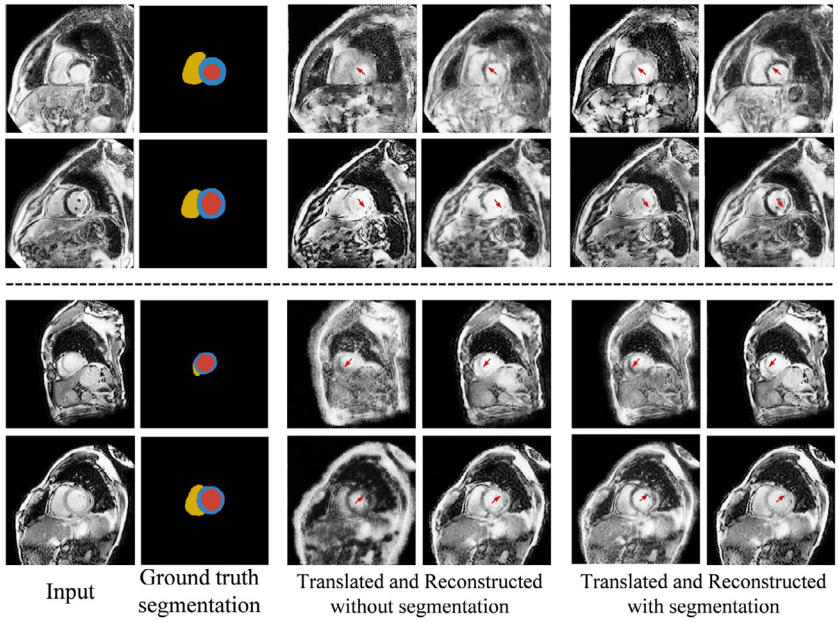

Fig. 7. Qualitative comparison of unsupervised image translation and reconstruction between our method with or without the semantic consistency loss incorporated. Top two rows: examples when the LGE CMR images are translated to the bSSFP CMR images; bottom two row: examples when the bSSFP CMR images are translated to the LGE CMR images.

in Table 4. From this table, one can see that the proposed CyCMIS method achieved an average DSC of 83.58% when evaluated on the MM-WHS challenge dataset, and an average DSC of 79.03% when evaluated on the MS-CMRSeg challenge dataset. Removing the diverse image augmentation block led to an average DSC drop of 3.30% and 5.85%, respectively, when evaluated on the two datasets. Further removing the intra-domain semantic consistency loss, a further drop of 2.92% and 3.34% average DSC was observed. Completely removing the semantic consistency loss (both inter-domain and intra-domain semantic consistency loss) resulted in a significant performance drop, i.e., the average DSC decreased 9.04% and 14.52%, respectively, when evaluated on the same two datasets. The qualitative results in Fig. 7 demonstrate how the incorporation of semantic consistency loss facilitate better image translation, which may explain why CyCMIS achieved better results than other SOTA methods. Both the qualitative and quantitative results demonstrate the effectiveness of different components of the proposed CyCMIS method on overcoming the domain shift when applied to both cross-modality and cross-protocol image adaptation.

Influence of different segmentation networks on the performance of CyCMIS. Using the same two datasets, we also investigated the influence of different segmentation networks on the performance of CyCMIS. Specifically, we replaced the PSPNet (Zhao et al., 2017) by the U-Net with attention block (referred as AB-U-Net), which was recently demonstrated by Wu et al. (2021a) to have better performance than the conventional U-Net (Ronneberger et al., 2015). We used DSC as the evaluation metric. The results are shown in Table 5. It can be seen that the PSPNet (Zhao et al., 2017) achieved better results than the AB-U-Net (Wu et al., 2021a).

5. Discussions

Developing UDA methods for cross-domain medical image segmentation is challenging. The goal of the present study is to develop and validate an accurate method that can mitigate such a challenge. In this paper, we presented a cycle-consistent cross-domain medical image segmentation method called CyCMIS, integrating online diverse image translation via disentangled representation learning and semantic consistency regularization into one network. We conducted extensive validation studies to evaluate the performance of the proposed CyCMIS model when it was applied to two typical yet challenging cross-modality/cross-protocol image segmentation tasks. Qualitative and quantitative results demonstrated that the proposed CyCMIS model was able to achieve results that were close or even better than the "Full supervision" upper bound. We showed in Fig. 3 that the disentangled content space provided a shared representation of structural information in different domains. We further investigated the influence of different components of CyCMIS on the segmentation performance. As shown in Table 4, different components such as diverse image augmentation, inter-domain and intra-domain semantic consistency loss all contributed to the improved performance of the proposed method. We additionally showed in Fig. 7 that the proposed semantic consistency loss could help to improve quality of both translated and cross-cycle recon-

Table 5
Quantitative results of the study investigating the influence of different segmentation networks on the performance of CyCMIS.

Segmentation Networks	MM-WHS challenge dataset			MS-CMR challenge dataset			
	LV	MYO	Average	LV	MYO	RV	Average
PSPNet	**90.82**	**76.33**	**83.58**	**87.11**	**70.75**	**79.24**	**79.03**
AB-U-Net	85.40	76.00	80.70	84.35	66.20	78.01	76.19

Table 6
A comparison of adaptation strategies adopted by the STOA unsupervised cross-domain segmentation methods.

Methods	Adaptation strategies
No adaptation	Training with source label only
AdaptSegNet (Tsai et al., 2018)	Feature alignment
ICMSC (Zeng et al., 2020)	Image appearance adaptation via CycleGAN
CyCADA (Hoffman et al., 2018)	Image appearance adaptation via CycleGAN + Feature alignment
SIFA (Chen et al., 2020a)	Image appearance adaptation via CycleGAN + Feature alignment
DDA-GAN (Chen et al., 2021)	Image appearance adaptation via disentangled representation learning
CyCMIS (ours)	Image appearance adaptation via disentangled representation learning

structed images, which might in turn benefit cross-domain image segmentation.

In comparison with the SOTA methods, the present approach achieved comparable or better results. On both datasets, we compared the results achieved by our method to those achieved by SIFA model introduced in (Chen et al., 2020a). For each dataset, both methods were evaluated using the same protocol. Thus, the results achieved by both methods could be directly compared. When evaluated on the MM-WHS challenge dataset, as presented in Table 1, our method achieved an average DSC of 84.51% when taking MR data as the target domain and an average DSC of 84.77% when taking CT data as the target domain, while their method achieved an average DSC of 81.14% and 81.04%, respectively. Furthermore, when evaluated on the MS-CMR challenge dataset, as presented in Table 2, their method obtained an average DSC of 74.88% when taking LGE CMR images as the target domain. In contrast, our method achieved an average DSC of 79.08%. Although both methods were designed for unsupervised cross-domain image segmentation, there are fundamental differences between two methods. First, as shown in Table 6, different image adaptation strategies were used. Specifically, SIFA (Chen et al., 2020a) combines feature alignment with image appearance adaptation via CycleGAN while our method conducts image appearance adaptation via disentangled representation learning, which, as shown in Fig. 6, allows for generating diverse images with unpaired training data. Second, their method leverages only one segmentation network in the target domain which is trained using only translated source domain images with labels. In contrast, as shown in Eq. (21) and (22), we train two segmentation networks, S_x in the source domain and S_y in the target domain. the segmentation network in the source domain S_x is well trained with supervised loss L_{sup} while the segmentation network in the target domain S_y is trained not only with the diversely translated images y'_i with labels m_x, which were generated by combining the content features of the source domain with multiple attribute codes of the target domain, but also the target domain training images y with pseudo-labels $S_x(x')$ generated by applying the well-trained S_x to the translated images x'. x' was generated by combining the content feature of the target domain with the attribute code of the source domain. Additionally, we also require that S_y, when applied to the cross-cycle reconstructed images $\hat{y}$, generates consistent segmentation results as the original images y, in order to further regularize the training of S_y.

The differences between the proposed CyCMIS method with DDA-GAN model introduced in (Chen et al., 2021) need to be discussed as both methods were based on disentangled representation learning and leveraged diverse data augmentation. First, the segmentation was performed in different spaces. Specifically, the method introduced in (Chen et al., 2021) performed cross-domain image segmentation on the disentangled content space while our proposed CyCMIS method performed segmentation on image space, taking attribute information into consideration. The diverse image augmentation block in our method combined disentangled content features with different attribute latent vectors to produce diverse images with the same content. It has been shown previously by Chen et al. (2019b) in a multi-stage framework that generating realistic and diverse synthetic images in the target domain given a single image in the source domain can enable generative model-based data augmentation for improving the generalizability of the segmentation network. This is exactly the motivation behind the design of our diverse data augmentation block. Different from the method presented in (Chen et al., 2019b), however, our method can be trained end-to-end. Second, the method introduced in (Chen et al., 2021) contained only one segmentation network which was trained using cross-domain semantic consistency enforcing the disentangled content features of the synthesized images after translation to be segmented exactly the same as before translation. In contrast, our method leveraged two segmentation networks, which were trained using both intra-domain and inter-domain (cross-domain) semantic consistency losses. Results in Table 4 showed that intra-domain semantic consistency loss could help to improve the cross-domain segmentation performance. Third, for the diverse data augmentation, the method introduced in (Chen et al., 2021) used a fixed number of 2 for each input image in the source domain while we conducted an experiment to determine the optimal value for the number of diversely translated images. Quantitative and qualitative results shown respectively in Table 1, 2 and Fig. 4, 5 demonstrated that the proposed CyCMIS model achieved better results than the DDS-GAN model when evaluated on two public datasets.

It is worth to discuss the limitations of the present study. First, the proposed CyCMIS, as a disentangled representation learning-based UDA method, is arguably complex than other SOTA cross-domain segmentation methods. However, most of the disentangled representation learning-based methods involve multiple losses which look complex but not so hard to train as exampled by sev-

eral previous works (Huang et al., 2018; Lee et al., 2020; Chen et al., 2021). Our experience showed that when we followed the training strategy as shown in Algorithm 1, the training process was stable and rarely failed. Second, the proposed CyCMIS method achieved relatively higher average DSC when evaluated on the MM-WHS challenge dataset than on the MS-CMRSeg challenge dataset, despite the fact that the former dataset showed relatively larger domain shift than the latter dataset. One possible explanation may be due to the fact that the background of the MM-WHS challenge dataset is relatively simpler than that of the MS-CMRSeg challenge dataset, as demonstrated by the disentangled content-only images shown in Fig. 3. Although we followed the practice introduced in (Wu and Zhuang, 2020) to prepare data, it is worth to investigate whether the same performance can be achieved when more complex background is involved. Third, when we designed experiments on the MS-CMRSeg challenge dataset, we took into consideration the fact that while the borders of the myocardium is difficult to delineate on LGE CMR images, they are clear and easy to identify on the bSSFP CMR images (Chen et al., 2021). Previous methods (Zhuang, 2018; Tao et al., 2015) use the segmentation results from the bSSFP CMR images of the same patient to assist the segmentation on LGE CMR images, which generally requires accurate registration between the bSSFP and LGE images. We thus only conducted the unsupervised adaptation from the bSSFP CMR images (the source domain) to the LGE CMR imges (the target domain), eliminating the requirement of paired bSSFP and LEG images and accurate registration, which was a clear advantage. In the future, We can also conduct an unsupervised adaptation from LGE CMR images to bSSFP CMR images but the practical value of such an adaptation needs to be investigated first.

6. Conclusion

In this paper, we proposed an end-to-end unsupervised cross-domain medical image segmentation method, taking advantage of diverse image translation via disentangled representation learning and consistency regularization into one network. We characterized the complex relationship between domains as many-to-many mapping and introduced a novel diverse inter-domain semantic consistency loss to regularize the cross-domain segmentation process. We additionally introduced an intra-domain semantic consistency loss to encourage the segmentation consistency between the original input and the image after cross-cycle reconstruction. We conducted comprehensive experiments on two publicly available datasets to evaluate the effectiveness of the proposed method. The experimental results demonstrated that the present method achieved better results than the SOTA methods. Our future work will focus on the scenario when more complex background is involved.

Declaration of Competing Interest

None.

CRediT authorship contribution statement

Runze Wang: Methodology, Software, Validation, Writing – original draft. **Guoyan Zheng:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Supervision, Funding acquisition.

Acknowledgments

This study was partially supported by Shanghai Municipal Science and Technology Commission via Project 20511105205 and by the Natural Science Foundation of China via project U20A20199.

References

Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.-A., Cetin, I., Lekadir, K., Camara, O., Ballester, M.A.G., et al., 2018. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE Trans Med Imaging 37 (11), 2514–2525.

Cai, J., Zhang, Z., Cui, L., Zheng, Y., Yang, L., 2019. Towards cross-modal organ translation and segmentation: a cycle-and shape-consistent generative adversarial network. Med Image Anal 52, 174–184.

Chartsias, A., Joyce, T., Dharmakumar, R., Tsaftaris, S.A., 2017. Adversarial Image Synthesis for Unpaired Multi-modal Cardiac Data. In: International workshop on simulation and synthesis in medical imaging. Springer. 3–13

Chartsias, A., Joyce, T., Papanastasiou, G., Semple, S., Williams, M., Newby, D.E., Dharmakumar, R., Tsaftaris, S.A., 2019. Disentangled representation learning in cardiac image analysis. Med Image Anal 58, 101535.

Chartsias, A., Papanastasiou, G., Wang, C., Semple, S., Newby, D.E., Dharmakumar, R., Tsaftaris, S.A., 2020. Disentangle, align and fuse for multimodal and semi-supervised image segmentation. IEEE Trans Med Imaging.

Chen, C., Dou, Q., Chen, H., Qin, J., Heng, P.A., 2020. Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation. IEEE Trans Med Imaging 39 (7), 2494–2505.

Chen, C., Dou, Q., Jin, Y., Chen, H., Qin, J., Heng, P.A., 2019. Robust Multimodal Brain Tumor Segmentation via Feature Disentanglement and Gated Fusion. In: In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer. Pp. 447–456

Chen, C., Ouyang, C., Tarroni, G., Schlemper, J., Qiu, H., Bai, W., Rueckert, D., 2019. Unsupervised Multi-modal Style Transfer for Cardiac Mr Segmentation. In: International Workshop on Statistical Atlases and Computational Models of the Heart. Springer, pp. 209–219.

Chen, X., Lian, C., Wang, L., Deng, H., Kuang, T., Fung, S., Gateno, J., Yap, P.-T., Xia, J.J., Shen, D., 2020. Anatomy-regularized representation learning for cross-modality medical image segmentation. IEEE Trans Med Imaging 40 (1), 274–285.

Chen, X., Lian, C., Wang, L., Deng, H., Kuang, T., Fung, S.H., Gateno, J., Shen, D., Xia, J.J., Yap, P.T., 2021. Diverse data augmentation for learning image segmentation with cross-modality annotations. Med Image Anal 71, 102060.

Cheplygina, V., Pena, I.P., Pedersen, J.H., Lynch, D.A., Sørensen, L., de Bruijne, M., 2017. Transfer learning for multicenter classification of chronic obstructive pulmonary disease. IEEE J Biomed Health Inform 22 (5), 1486–1496.

Choudhary, A., Tong, L., Zhu, Y., Wang, M.D., 2020. Advancing medical imaging informatics by deep learning-based domain adaptation. Yearb Med Inform 29 (1), 129.

Dou, Q., Ouyang, C., Chen, C., Chen, H., Glocker, B., Zhuang, X., Heng, P.A., 2019. Pnp-adanet: plug-and-play adversarial domain adaptation network at unpaired cross-modality cardiac segmentation. IEEE Access 7, 99065–99076.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V., 2016. Domain-adversarial training of neural networks. The journal of machine learning research 17 (1). 2096–2030

Guan, H., Liu, M., 2021. Domain adaptation for medical image analysis: a survey. arXiv preprint arXiv:2102.09508.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep Residual Learning for Image Recognition. In: In: Proceedings of the IEEE conference on computer vision and pattern recognition pp. 770–778.

Heimann, T., Mountney, P., John, M., Ionasec, R., 2013. Learning without Labeling: Domain Adaptation for Ultrasound Transducer Localization. In: In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 49–56.

Hiasa, Y., Otake, Y., Takao, M., Matsuoka, T., Takashima, K., Carass, A., Prince, J.L., Sugano, N., Sato, Y., 2018. Cross-modality Image Synthesis from Unpaired Data Using Cyclegan. In: International workshop on simulation and synthesis in medical imaging. Springer, pp. 31–41.

Hoffman, J., Tzeng, E., Park, T., Zhu, J.-Y., Isola, P., Saenko, K., Efros, A., Darrell, T., 2018. Cycada: Cycle-consistent Adversarial Domain Adaptation. In: International conference on machine learning. PMLR, pp. 1989–1998.

Huang, X., Liu, M.-Y., Belongie, S., Kautz, J., 2018. Multimodal unsupervised image–to-image translation. Proceedings of the European conference on computer vision (ECCV) 172–189.

Huo, Y., Xu, Z., Moon, H., Bao, S., Assad, A., Moyo, T.K., Savona, M.R., Abramson, R.G., Landman, B.A., 2018. Synseg-net: synthetic segmentation without target modality ground truth. IEEE Trans Med Imaging 38 (4), 1016–1025.

Jiang, J., Hu, Y.-C., Tyagi, N., Zhang, P., Rimner, A., Mageras, G.S., Deasy, J.O., Veeraraghavan, H., 2018. Tumor-aware, Adversarial Domain Adaptation from Ct to Mri for Lung Cancer Segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 777–785.

Lee, H.-Y., Tseng, H.-Y., Mao, Q., Huang, J.-B., Lu, Y.-D., Singh, M., Yang, M.H., 2020. Drit++: diverse image-to-image translation via disentangled representations. Int J Comput Vis 128 (10), 2402–2417.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Med Image Anal 42, 60–88.

Liu, F., 2019. Susan: segment unannotated image structure using adversarial network. Magn Reson Med 81 (5), 3330–3345.

Mao, Q., Lee, H.-Y., Tseng, H.-Y., Ma, S., Yang, M.H., 2019. Mode Seeking Generative Adversarial Networks for Diverse Image Synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1429–1437.

Patravali, J., Jain, S., Chilamkurthy, S., 2017. 2D-3D Fully Convolutional Neural Networks for Cardiac Mr Segmentation. In: International Workshop on Statistical Atlases and Computational Models of the Heart. Springer, pp. 130–139.
Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional Networks for Biomedical Image Segmentation. In: In: International Conference on Medical image computing and computer-assisted intervention. Springer, pp. 234–241.
Tao, Q., Piers, S.R., Lamb, H.J., van derGeest, R.J., 2015. Automated left ventricle segmentation in late gadolinium-enhanced mri for objective myocardial scar assessment. J. Magn. Reson. Imaging 42 (2), 390–399.
Tsai, Y.-H., Hung, W.-C., Schulter, S., Sohn, K., Yang, M.-H., Chandraker, M., 2018. Learning to Adapt Structured Output Space for Semantic Segmentation. In: In: Proceedings of the IEEE conference on computer vision and pattern recognition pp. 7472–7481.
Wu, F., Zhuang, X., 2020. Cf distance: a new domain discrepancy metric and application to explicit domain adaptation for cross-modality cardiac image segmentation. IEEE Trans Med Imaging 39 (12), 4274–4285.
Wu, Y., Hatipoglu, S., Alonso-Álvarez, D., Gatehouse, P., Li, B., Gao, Y., Firmin, D., Keegan, J., Yang, G., 2021. Fast and automated segmentation for the three-directional multi-slice cine myocardial velocity mapping. Diagnostics 11 (2), 346.
Wu, Y., Tang, Z., Li, B., Firmin, D., Yang, G., 2021. Recent advances in fibrosis and scar segmentation from cardiac mri: a state-of-the-art review and future perspectives. Front Physiol 12.
Yan, W., Wang, Y., Gu, S., Huang, L., Yan, F., Xia, L., Tao, Q., 2019. The Domain Shift Problem of Medical Image Segmentation and Vendor-adaptation by Unet-gan. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 623–631.
Yan, W., Wang, Y., Xia, M., Tao, Q., 2019. Edge-guided output adaptor: highly efficient adaptation module for cross-vendor medical image segmentation. IEEE Signal Process Lett 26 (11), 1593–1597.
Yang, D., Hong, S., Jang, Y., Zhao, T., Lee, H., 2019. Diversity-sensitive conditional generative adversarial networks. arXiv preprint arXiv:1901.09024.
Yang, G., Ye, Q., Xia, J., 2022. Unbox the black-box for the medical explainable ai via multi-modal and multi-centre data fusion: a mini-review, two showcases and beyond. Information Fusion 77, 29–52.
Yang, G., Zhang, H., Firmin, D., Li, S., 2021. Recent advances in artificial intelligence for cardiac imaging. Computerized medical imaging and graphics 90, 101928.
Yang, J., Dvornek, N.C., Zhang, F., Chapiro, J., Lin, M., Duncan, J.S., 2019. Unsupervised Domain Adaptation via Disentangled Representations: Application to Cross-modality Liver Segmentation. In: In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 255–263.
Zeng, G., Lerch, T.D., Schmaranzer, F., Zheng, G., Burger, J., Gerber, K., Tannast, M., Siebenrock, K., Gerber, N., 2020. Icmsc: intra-and cross-modality semantic consistency for unsupervised domain adaptation on hip joint bone segmentation. arXiv preprint arXiv:2012.12570.
Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid Scene Parsing Network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2881–2890.
Zhu, J.-Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired Image-to-image Translation Using Cycle-consistent Adversarial Networks. In: In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232.
Zhuang, X., 2018. Multivariate Mixture Model for Myocardial Segmentation Combining Multi-source Images. In: IEEE transactions on pattern analysis and machine intelligence, Vol. 41, pp. 2933–2946.
Zhuang, X., Li, L., Payer, C., Štern, D., Urschler, M., Heinrich, M.P., Oster, J., Wang, C., Smedby, O., Bian, C., et al., 2019. Evaluation of algorithms for multi-modality whole heart segmentation: an open-access grand challenge. Med Image Anal 58, 101537.