Medical Image Analysis 88 (2023) 102873

Contents lists available at ScienceDirect

Medical Image Analysis

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

A novel one-to-multiple unsupervised domain adaptation framework for abdominal organ segmentation

Xiaowei Xu [a,b], Yinan Chen [a,b,d,*], Jianghao Wu [b], Jiangshan Lu [b], Yuxiang Ye [a], Yechong Huang [a], Xin Dou [f], Kang Li [c,d,e], Guotai Wang [b,c], Shaoting Zhang [a,b,c], Wei Gong [g,h]

[a] SenseTime Research, Shanghai, China
[b] School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu, Sichuan, China
[c] Shanghai Artificial Intelligence Laboratory, Shanghai, China
[d] West China Biomedical Big Data Center, West China Hospital, Sichuan University, Chengdu, Sichuan, China
[e] Med-X Center for Informatics, Sichuan University, Chengdu, Sichuan, China
[f] SenseBrain Technology, Princeton, NJ 08540, USA
[g] Department of General Surgery, Xinhua Hospital, Affiliated to Shanghai Jiao Tong University School of Medicine, Shanghai, 200092, China
[h] Shanghai Key Laboratory of Biliary Tract Disease Research, Shanghai, 200092, China

ARTICLE INFO

MSC:
41A05
41A10
65D05
65D17

Keywords:
Abdominal multi-organ segmentation
Multi-modality MRI
One-to-multiple unsupervised domain adaption
Image translation

ABSTRACT

Abdominal multi-organ segmentation in multi-sequence magnetic resonance images (MRI) is of great significance in many clinical scenarios, e.g., MRI-oriented pre-operative treatment planning. Labeling multiple organs on a single MR sequence is a time-consuming and labor-intensive task, let alone manual labeling on multiple MR sequences. Training a model by one sequence and generalizing it to other domains is one way to reduce the burden of manual annotation, but the existence of domain gap often leads to poor generalization performance of such methods. Image translation-based unsupervised domain adaptation (UDA) is a common way to address this domain gap issue. However, existing methods focus less on keeping anatomical consistency and are limited by one-to-one domain adaptation, leading to low efficiency for adapting a model to multiple target domains. This work proposes a unified framework called OMUDA for one-to-multiple unsupervised domain-adaptive segmentation, where disentanglement between content and style is used to efficiently translate a source domain image into multiple target domains. Moreover, generator refactoring and style constraint are conducted in OMUDA for better maintaining cross-modality structural consistency and reducing domain aliasing. The average Dice Similarity Coefficients (DSCs) of OMUDA for multiple sequences and organs on the in-house test set, the AMOS22 dataset and the CHAOS dataset are 85.51%, 82.66% and 91.38%, respectively, which are slightly lower than those of CycleGAN(85.66% and 83.40%) in the first two data sets and slightly higher than CycleGAN(91.36%) in the last dataset. But compared with CycleGAN, OMUDA reduces floating-point calculations by about 87 percent in the training phase and about 30 percent in the inference stage respectively. The quantitative results in both segmentation performance and training efficiency demonstrate the usability of OMUDA in some practical scenes, such as the initial phase of product development.

1. Introduction

Organ segmentation on medical images has become an important task for many clinical applications, such as computer-aided diagnosis, surgical planning, and radiotherapy. In the past 10 years, a series of huge progress have been made in abdominal organ segmentation based on convolutional neural network (CNN) (Ma et al., 2021; Kavur et al., 2021).

Early works (Karasawa et al., 2017; Dou et al., 2016) on abdominal organ segmentation mainly focused on the segmentation of single organs in computed tomography (CT) images. Recently, due to the benefit of image acquisition without using ionizing radiation and the better contrast of soft tissue on the abdomen, magnetic resonance imaging (MRI) gradually becomes a more attractive modality for daily imaging throughout diagnosis, treatment and follow-ups. Abdominal multi-organ segmentation from MR images plays an important role in many clinical scenarios. For instance, in MRI-based radiotherapy (Savenije et al., 2020) which has become a current focus of clinical interest, accurate location of the target organ and precise delineation of neighboring organs can help minimize normal tissue toxicity. Moreover, segmentation of multiple organs on MRI is also a prerequisite for

* Corresponding author at: SenseTime Research, Shanghai, China.
E-mail address: chenyinannan@hotmail.com (Y. Chen).

https://doi.org/10.1016/j.media.2023.102873
Received 7 November 2022; Received in revised form 24 March 2023; Accepted 12 June 2023
Available online 16 June 2023

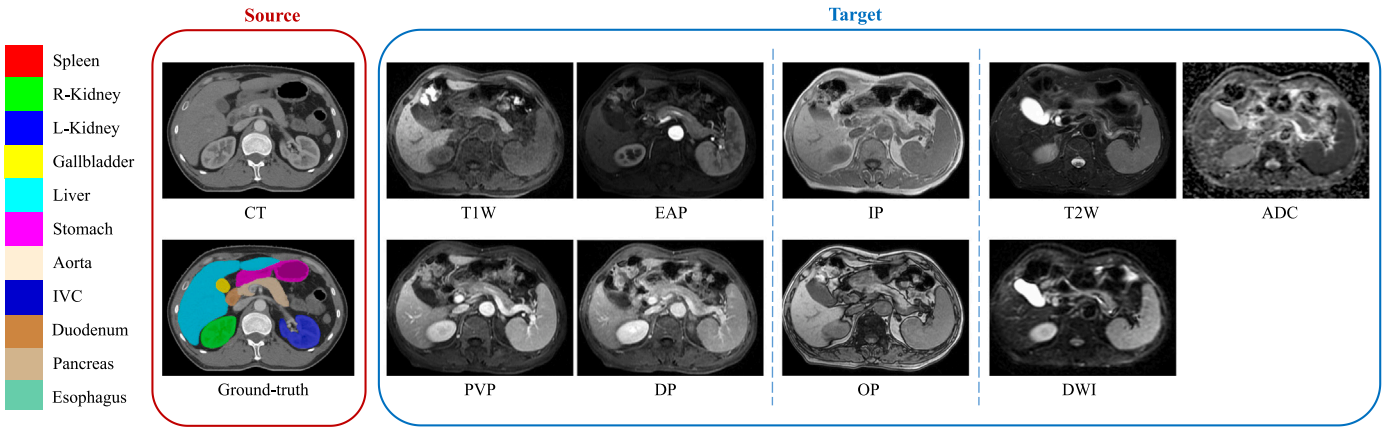


Fig. 1. Illustration of the task in this work. We aim to segment 11 abdominal organs in 9 different MR sequences by learning from existing annotated CT datasets without annotations in the target domains.

the detection of different types of primary or metastatic tumors across the entire abdominal organs.

In the vast majority of cases, a single MRI sequence cannot meet clinical needs. As shown in Fig. 1, different MRI sequences such as T1-weighted (T1 W), T2-weighted (T2 W), dynamic-contrast enhancement, diffusion-weighted image (DWI), and in-phase (IP)/out-of-phase (OP), have different organization-contrast mechanisms, and can provide complementary information about anatomy and tumor's physiological and metabolic function. For example, T1 W and T2 W can be used for the diagnosis of liver cirrhosis and bleeding, but may not be suitable for the diagnosis of benign and malignant lesions. Dynamic-contrast enhanced sequences can make up for the deficiency of non-contrast enhanced sequences in the diagnosis of benign and malignant lesions. Therefore, in clinical practice, different patients may be scanned with different MR sequences based on the different diseases and clinician's preference, and it is desirable to segment the abdominal organs in a range of MR sequences.

However, current fully supervised paradigm can hardly be applied to such a scenario (Duan et al., 2020), as they are not scalable due to the requirement of annotating a large set of images in each target sequence for training. Annotating all the possible target sequences respectively is extremely time-consuming and inefficient with repeated works. As medical images in different sequences share the same set of anatomical structures with different appearances, it is desirable to only annotate these structures in a single source modality/sequence to train models to segment images from a range of target sequences. Computer-aided abdominal multi-organ segmentation in CT images has been relatively mature, and more and more public CT abdominal databases are accessible. To reduce the annotation cost and avoid repeated annotation in different MR sequences, it is desirable to leverage these annotated CT images to train deep learning models to segment the MR sequences. As shown in Fig. 1, this work aims to segment 11 abdominal organs from multiple MR sequences respectively by learning from existing annotated CT datasets without annotations in the target MR sequences.

The above problem can be attributed to the unsupervised domain adaptation (UDA) problem (Wu et al., 2022). For the UDA problem with a large domain gap, the most commonly used method is to translate the image based on generative adversarial networks (GANs) to achieve the adaptation of the image appearance. But a vast majority of generator in existing methods can only complete the translation from one source domain to a single target domain, which cannot be scaled to one-to-multiple translation tasks due to the large amount of computation and time consumption of using one-to-one translation for multiple times. Frameworks such as StarGAN (Choi et al., 2018) and StarGAN v2 (Choi et al., 2020) have made some attempts in one-to-multiple domain image translation and achieved satisfactory results in the field of natural image synthesis. However, these frameworks concentrate more on the realistic appearance of synthetic images and less on the cross-domain structural consistency which is critical for medical images. Consequently, these methods may cause distortion and deformation of the anatomical structure, which will limit the segmentation accuracy on the target domain images.

In this paper, we propose a novel One-to-Multiple Unsupervised Domain Adaptation (OMUDA) framework to automatically segment 11 abdominal organs, i.e., spleen, right kidney (R-Kidney), left kidney (L-Kidney), gallbladder, liver, stomach, aorta, interior vena cava (IVC), duodenum, pancreas and esophagus, in 9 sequences of MRI images, i.e., T1 W, early-arterial phase (EAP), portal venous phase (PVP), delay phase (DP), IP, OP, T2 W, DWI and apparent diffusion coefficient map (ADC), using existing annotated CT datasets. As shown in 2, OMUDA uses disentanglement between content and style features to efficiently translate a source domain image into multiple target domains with a single generator. In addition, generator refactoring and style constraint are implemented to guarantee the consistency of anatomical structure during image translation and reduce domain aliasing. Extensive experiments indicate its advantages in both segmentation performance and training efficiency.

The main contribution of our work lies in the following aspects:

1. We proposed a unified framework for one-to-multiple unsupervised domain adaptation in segmentation of abdominal organs, where a model trained from a single source domain is adapted to multiple target domains simultaneously and efficiently without extra annotations.
2. Our method drew inspiration from the generator architecture of CycleGAN and the multi-domain inter-transformation concept of StarGAN v2 to achieve efficient image translation between arbitrary pairs of domains with a single generator, as well as for better cross-modality structural consistency.
3. Style constraint was explicitly imposed on the generated style codes to minimize the inter-domain style variance and maximize the intra-domain style variance, and thus reduce domain aliasing.
4. Our method achieved accurate segmentation results (an average Dice Similarity Coefficient (DSC) of 85.51% on the in-house test set, an average Dice of 82.66% and 91.38% on two external datasets) when simultaneously adapting a model from CT to 9 MR sequences, avoiding time-consuming adaptation to each target domain respectively.

2. Related work

2.1. Abdominal organ segmentation

Abdominal organ segmentation is an important clinical task. The shape, size, and direction of abdominal organs vary greatly, which

brings great challenges to the segmentation task. Early abdominal organ segmentation studies usually focused on a single organ. Graphical (Wu et al., 2016) and deformable (Chartrand et al., 2016) models were proposed for automatic liver segmentation. Roth et al. (2018) applied a bigger patch size to deal with the whole dense pancreatic volume. These works are suitable for the segmentation of a single organ, however, their poor scalability is not conducive to the simultaneous segmentation of multiple organs. For multi-organs segmentation, many works adopt atlas-based (Karasawa et al., 2017) methods, but the computational time and segmentation accuracy still need to be improved. Recently, researchers have demonstrated that fully convolutional neural networks show great promise for abdominal organ segmentation in CT scans. Zhou et al. (2016) segmented abdominal multi-organs on 2D slices in axial, sagittal, and coronal planes, and fused the combined segmentation results using majority voting labels. Roth et al. (2017) proposed a coarse-to-fine two-stage approach for abdominal multi-organ segmentation, where the first stage roughly delineates the organ of interest and the second stage focuses on more detailed organ segmentation. Gibson et al. (2018) proposed a deep learning-based registration-free segmentation algorithm for the segmentation of eight abdominal organs.

However, a single modality is often difficult to meet the clinical needs because of the limited amount of information it can provide. In clinical application, combining multi-modal images for diagnosis and treatment has gradually become a mainstream trend, which leads to necessity of abdominal organ segmentation on multi-modal images. Bobo et al. (2018) and Kart et al. (2021) utilized fully convolutional neural networks for abdominal multi-organ segmentation in T2 W and T1 W sequences respectively. During the CHAOS challenge (Kavur et al., 2021), researchers explored the task of multi-modal abdominal organ segmentation. For example, Conze et al. (2021) proposed a model based on Conditional Generative Adversarial Networks (cGANs) for abdominal organs segmentation in IP, OP and T2 W, which embeds cascaded partially pre-trained convolutional encoder–decoders as generator to learn how to delineate organs and a discriminator to enforce the model to create realistic segmentation masks. They won the first prize in both Task 3 and Task 5 of the CHAOS challenge. Pham et al. (2019) incorporated anatomical priors to the segmentation network for helping the abdominal organs segmentation in CT and MRI, and got the first place in Task 4 of the CHAOS challenge. Isensee et al. (2019b) utilized an internal variant of nnU-Net and some tricks of ensemble learning for segmenting the abdominal organs in IP, OP and T2 W. However, the methods are designed for specific sequences in a fully supervised paradigm, which have limited scalability for the multi-sequence segmentation task due to the high annotation cost.

2.2. Unsupervised domain adaptation

Unsupervised Domain Adaptation (UDA) aims to transfer knowledge from a labeled source domain to an unlabeled target domain. Existing works have achieved impressive results for one-to-one domain adaptation, and they can be roughly categorized into three types: (1) feature-based methods; (2) image translation-based methods; (3) pseudo-label-based methods.

Feature-based methods are designed to learn both semantically meaningful and domain-invariant features, and reduce the representation differences between domains, such as distributional differences and geometrical differences. Tzeng et al. (2014) minimize the domain gap by minimizing the domain confusion loss. Du et al. (2019) proposed a domain adaptation network based on separated semantic features, which addresses the inconsistent adaptation problem in class-wise adversarial learning. Unlike feature-based methods, image translation-based methods aim to learn a mapping between the source and target domains, and achieve the adaptation from the perspective of the image appearance. For example, CyCADA (Hoffman et al., 2018) uses cycle-consistent constraints to translate images through CycleGAN (Zhu et al., 2017) for domain adaptation. UNIT (Liu et al., 2017) combines a variational autoencoder and CoGAN (Liu and Tuzel, 2016) to map corresponding images in both domains to the same latent code. Huang et al. (2018) proposed a multimodal unsupervised image-to-image translation (MUNIT) framework which explicitly disentangle the image representation into content and style. In this framework, the domain-invariant content is combined with the domain-specific style to achieve image appearance adaptation. In regard to pseudo-label-based methods, Jiang et al. (2020) proposed a sampling-based implicit alignment method, in which sample selection is implicitly guided by pseudo-labels, which has better results under the situation of domain class imbalance. To mitigate the effect of noisy pseudo-labels, Ge et al. (2020) proposed a mean-teaching method to softly refine the pseudo labels in the target domain.

However, these one-to-one domain adaptation methods can only learn the relationship between two different domains, and has limited scalability when dealing with multiple domains. StarGAN (Choi et al., 2018) implemented a scalable image-to-image translation model across multiple domains by using a single generator and a discriminator. StarGAN v2 (Choi et al., 2020) further increased the diversity of generated images by introducing domain-specific style code to the generator. MM-GAN (Sharma and Hamarneh, 2019) as a multi-modal generative adversarial network which exhibited the ability to synthesize the required MRI sequences, and achieved brain segmentation under different MRI sequences. Gholami et al. (2020) learned more robust feature representations for each target domain by exploiting the dependencies between multiple target domains. Isobe et al. (2021) proposed a collaborative learning framework that fully explored the connection between each source-target domain pair and among target domains to achieve unsupervised multi-target domain adaptation. DRIT++ (Lee et al., 2020) introduced a disentangled representation framework for multi-domain image translation. However, these frameworks focus more on the fidelity of pseudo images but less on the structural consistency before and after image translation.

3. Method

In this paper, a novel one-to-multiple unsupervised domain adaptation (OMUDA) framework composed by a One-to-Multiple Domain Generation (OMDG) framework and a segmentation network is proposed to efficiently train deep learning models to segment multiple organs from a range of MR sequences by leveraging annotations of some existing annotated CT images, thereby avoiding the high cost of annotating each of the target MR sequences. The whole framework is illustrated in Fig. 2, and it consists of two stages: CT-to-Multi-sequence MRI translation and training with the translated target modality images. The former aims to train a generator which translates a CT image to its corresponding MR images with multiple sequences simultaneously, and the latter employs the generated multi-sequence MR images and the annotations of their corresponding source CT images to train segmentation networks.

3.1. CT-to-multi-sequence MRI translation

Three properties are expected for effective image translation from CT to multiple MR sequences: efficiency, geometry preservation and minimized domain gap between synthesized and real MR images. First, one-to-one image translation is inefficient for translating a source domain image to images in multiple target domains. For instance, if we want to translate CT to N MRI modalities, N one-to-one generators (e.g. CycleGAN (Zhu et al., 2017)) are supposed to be trained. Moreover, if mutual translation of these domains are expected, the number of generators reaches the level of $O(N^2)$. Thus in this paper, a generation framework based on StarGAN v2 (Choi et al., 2020) is proposed to handle the one-to-multiple domain image translation

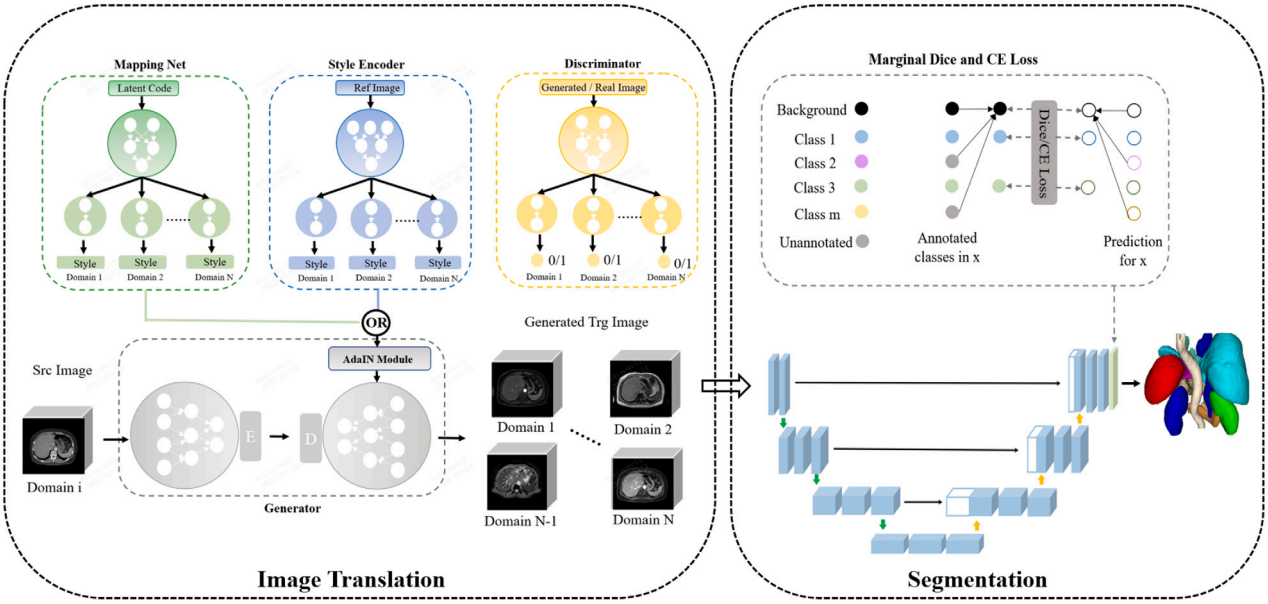


Fig. 2. The one-to-multiple unsupervised domain adaptation (OMUDA) framework.

problem and make the training and inference more efficient. Secondly, due to the property of the subsequent task that combines translated images with the organ masks of their corresponding source CT images to train segmentation networks, the structural consistency of organs needs to be ensured during the image translation. Therefore, we refactored the architecture of the generator in the StarGAN v2 to reduce the loss of organ structure information. Thirdly, in order to guarantee the segmentation performance of the network trained by pseudo images on real images, the generated images should have similar styles of their corresponding target modalities. In our generation framework, style constraint is utilized in the style encoder to help efficiently extract the target modality information.

One-to-Multiple Domain Generation Inspired by StarGAN v2, the OMDG framework was proposed to complete image translation between CT and arbitrary MR sequences with only one generator. As shown in Fig. 3), it consists of four networks: style encoder, mapping network, generator and discriminator. The style encoder extracts the style codes of reference images, which will be fed into the generator to implement reference-guided image translation. The mapping network also acts as a style code provider, which uses fully connected layers to map a latent code sampled from a Gaussian distribution into a style code. The generator translates an input image into an output image reflecting the domain-specific style code which is obtained from the style encoder or mapping network. The discriminator distinguishes between the real images and the generated images for each domain. To reduce the domain aliasing, the style encoder, mapping network and discriminator contain multiple output branches, and each branch servers for one certain target domain. Taking the discriminator as an example, each output branch is corresponding to one target domain and judges the real and fake images only for this domain. In our experiment, the architectures of the style encoder, mapping network and discriminator are kept same as those in the StarGAN v2.

Generator Architecture As mentioned above, the generated MR images and the annotation in the CT images will be paired to train the subsequent segmentation networks. Consequently, the organ structures are supposed be consistent before and after image translation. However, because of excessive downsampling layers and the lost spatial information caused by these downsampling operations, the original generator of StarGAN v2 is hard to keep the structural consistency of organs during the image translation (Fig. 7). So we refactored the architecture to reduce the loss of spatial information in the encoding phase of the generator. For the sake of the good performance of CycleGAN in one-to-one medical image translation, the refactored generator follows the architecture of its generator, but in order to realize one-to-multiple domain translation by one generator, the instance normalization (IN) layers in the decoder are replaced by the domain-specific adaptive instance normalization (AdaIN (Huang and Belongie, 2017)) layers, and more details about the architecture are shown in Table 3. The equation of the AdaIN layer is shown below.

$$AdaIN(f, s_t) = \sigma_t^l\left(\frac{f^l - \mu(f^l)}{\sigma(f^l)}\right) + \mu_t^l \tag{1}$$

$$(\mu_t^l, \sigma_t^l) = W^l \cdot s_t + b^l \tag{2}$$

where $f^l \in \mathbb{R}^{N\times C\times H\times W}$ and $s_t \in \mathbb{R}^{N\times S}$ are the feature map of the lth layer and the target style code respectively. $\mu(f^l) \in \mathbb{R}^{N\times C}$ and $\sigma(f^l) \in \mathbb{R}^{N\times C}$ are the mean and standard variance of f^l. $W^l \in \mathbb{R}^{S\times 2C}$ and $b^l \in \mathbb{R}^{2C}$ are the trainable weights and biases which are used to map the domain-specific styles to re-scale parameters of the lth layer, i.e., μ_t^l and σ_t^l. These layers receive the domain-specific style code generated by the style encoder or the mapping network, and transform it to layer-specific instance normalization parameters by fully connected layers. Then these parameters are used to re-scale the feature maps, and thus the domain-specific information is added into the feature maps automatically. With guidance of these domain-specific information, the generator gradually translates the source images to these target domains.

Style Constraint For reducing domain aliasing, multiple output branches are employed in the style encoder of StarGAN v2, however, no explicit constraint is set on the style codes of different domains. Metric learning losses are widely used to minimize the inter-class variance and maximize the intra-class variance. Inspired by N-pair loss (Sohn, 2016), a commonly used metric learning loss, a style-constrained loss is introduced into the training phase to add explicit constraints on the style codes of different domains. Similar to the contrastive learning, in a mini-batch, each style code is an anchor, and the style codes from the same domains as the anchor is deemed as the positives and vice versa. This style-constrained loss aims to decrease the difference between positive anchor pairs and increase the difference between negative anchor pairs, i.e., enlarging the distribution difference of style codes of different domains. The formula of this style-constrained loss is shown

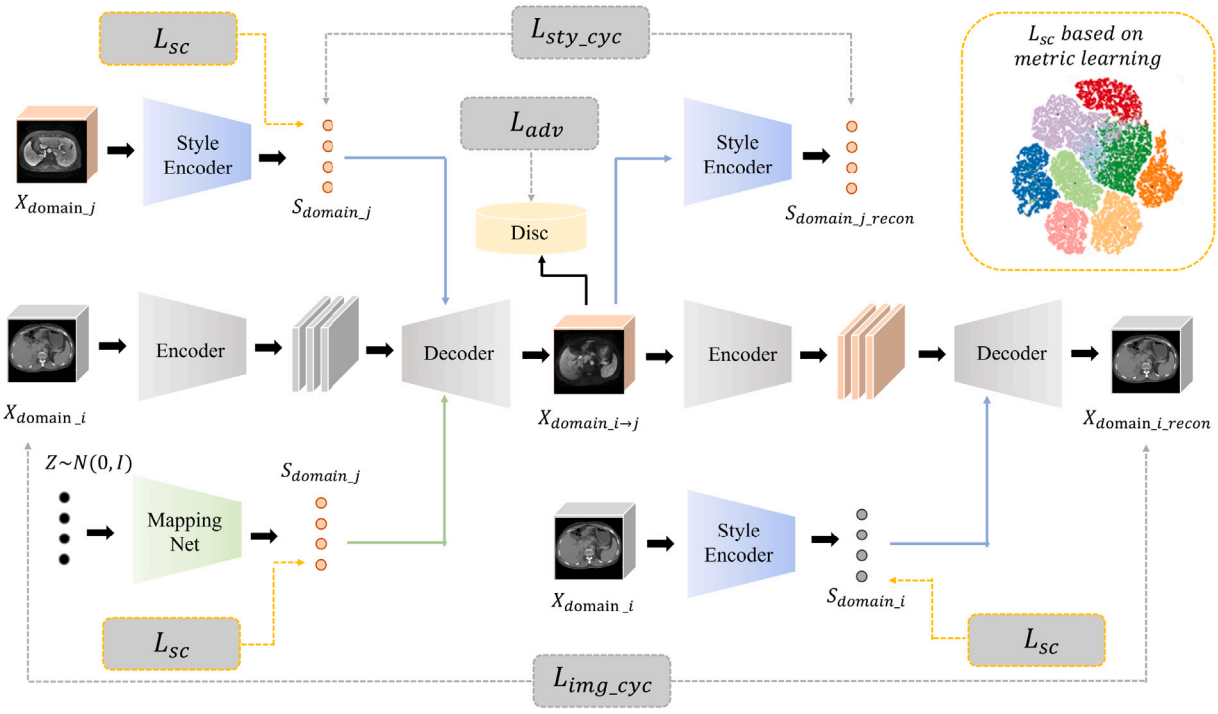


Fig. 3. Overview of the OMDG framework consisting of four networks. The style encoder and mapping network are style code providers, where the former extracts the style code from an image and the latter transforms a latent code into style codes for multiple domains. The generator translates a source image into a target image conditioned on a given domain-specific style code. The discriminator distinguishes between the real images and the generated images for each domain. Note that the style encoder, mapping network and discriminator contain multiple output branches as shown in Fig. 2.

below (Eq. (3)).

$$L_{\mathrm{sc}} = -\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|\Omega_i^+|}\sum_{s^+\in\Omega_i^+}\ln\frac{e^{s_i\cdot s^+/\tau}}{\sum_{j=1}^{N}e^{s_i\cdot s_j/\tau}} \tag{3}$$

where s_i is the style code extracted from the ith reference image in a mini-batch. Ω_i^+ is a set which is composed of the positives of anchor s_i. τ is the temperature, a hyper-parameter to scale the sensitivity of the loss function, and is set to 1 in our experiment.

Training Objectives Besides the style-constrained loss, the other loss functions for training the OMDG framework are the same as those in StarGAn v2. As mentioned above, the style codes can be provided by the mapping network or the style encoder. Taking the latent-mapping image translation as an example to illustrate the overall objectives. Give a source image $x_s \in \mathcal{X}$, its corresponding source domain $y_s \in \mathcal{Y} \subseteq \mathbb{N}^1$, target domain $y_t \in \mathcal{Y}$ and a latent vector $z \in N(0, I)$, we train the OMDG framework by the following objectives.

1. Adversarial Loss

$$L_{adv} = \mathbb{E}_{x_s,y_t,z}[\log(1 - D(G(x_s, M(z, y_t)), y_t))] + \mathbb{E}_{x_s}[\log D(x_s, y_s)] \tag{4}$$

where G, D and M denote the generator, discriminator and mapping network respectively.

2. Cycle-Consistency Loss

$$L_{img\_cyc} = \mathbb{E}_{x_s,y_s,y_t,z}[\|G(G(x_s, M(z, y_t)), E_s(x_s, y_s)) - x_s\|_1] \tag{5}$$

$$L_{sty\_cyc} = \mathbb{E}_{x_s,y_t,z}[\|E_s(G(x_s, M(z, y_t)), y_t) - M(z, y_t)\|_1] \tag{6}$$

where E_s denotes the style encoder. Cycle-consistency losses are composed of the image-cycle loss and the style-cycle loss. The former (L_{img_cyc}) encourages the generator G to reconstruct the input image x_s with its own style code $E_s(x_s, y_s)$, and thus enforces G to preserve the structural characteristics of x_s while changing its style faithfully. The latter (L_{sty_cyc}) enforces G to utilize the style code $M(z, y_t)$ when generating the image $G(xs, M(z, yt))$ for target domain.

3. Style Diversification Loss

$$L_{sd} = -\mathbb{E}_{x_s,z1,z2,y_t}[\|G(x_s, M(z1, y_t)) - G(x_s, M(z2, y_t))\|_1] \tag{7}$$

where $z1$, $z2$ are two latent vectors sampled from $N(0, I)$. This loss aims to increase the diversity of generated images, and thus avoid that a source image can only be translated to a specific image for each domain no matter how the input style changes.

Therefore, the full training loss an be summarized as:

$$\min_{G,M,E_s}\max_{D} L_{adv} + \lambda_{img\_cyc}L_{img\_cyc} + \lambda_{sty\_cyc}L_{sty\_cyc} + \lambda_{sd}L_{sd} + \lambda_{sc}L_{sc} \tag{8}$$

where λ_{img_cyc}, λ_{sty_cyc}, λ_{sd} and λ_{sc} are weights for each term. In our experiment, except λ_{sd}, they are all set to 1. λ_{sd} is initially set to 1 and linearly decreases to 0 with the number of iterations, which can ensure the fidelity of the generated images on the basic of diversity. Furthermore, in the same iteration, this OMDG framework is trained in the same manner as the above objective, using reference images instead of latent vectors when generating style codes.

Fake MRI generation Once the OMDG framework is trained, the generator combined with the style encoder or the mapping network can transform CT images to multi-sequence MR images. In this stage, so as to ensure the consistency of styles within single generated volume and the diversity of styles between different generated volumes, slices of the same generated MR volume share the same style code, while slices of different generated MR volumes differ in style codes. Similar as in the training phase, the style code can be provided either by the style encoder or by the mapping network in the inference stage.

Table 1
The annotation of 11 target organs in six public CT datasets used for training in this study. ✓ and ✗ mean the organ is labelled and unlabelled respectively.

Database	Spleen	R-Kidney	L-Kidney	Gallbladder	Liver	Stomach	Aorta	IVC	Duodenum	Pancreas	Esophagus
BTCV	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
TCIA	✓	✗	✓	✓	✓	✓	✗	✗	✓	✓	✓
NAFLD	✓	✓	✓	✗	✓	✗	✗	✗	✗	✗	✗
3D-IRCADb-01	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗
Decathlon	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗
WORD	✓	✓	✓	✓	✓	✓	✗	✗	✓	✓	✓

3.2. Training segmentation models in the target domains

After image translation, the generated MR images of multiple sequences and the masks of their corresponding source CT images are paired to train the segmentation networks, each dealing with a specific MR sequence. The training process is annotation-free for the target MR sequences when the source CT images are annotated. Nevertheless, it is time-consuming and label-intensive to delineate the abdominal multi-organ masks on CT images. Fortunately, some public dataset, e.g., Beyond the Cranial Vault (BTCV), the Cancer Imaging Archive (TCIA), have provided CT images with annotations of multiple organs. However, different datasets have different sets of annotated organs. For fully utilizing the samples from different datasets and handling this partial label learning problem, marginal Dice loss and cross entropy (CE) loss (Shi et al., 2020) are used to train the segmentation networks in our study. The key of marginal loss is the label or prediction recalibration (Marginal Dice Fig. 2). Given the network output probability p, its corresponding annotated mask g which is represented by a one-hot multi-channel image, and the indication vector $\alpha \in \{0,1\}$ indicating whether an organ is annotated or not, the label and prediction recalibration is defined as follows where classes without annotations are grouped into the background class, and the corresponding predicted probabilities of these classes are also superimposed on the predicted probabilities of the background class.

$$p'_j = \sum_{i \in \Omega_j} p_i \tag{9}$$

$$g'_j = \sum_{i \in \Omega_j} g_i \tag{10}$$

$$\Omega_j = \begin{cases} \{i|\alpha_i = 0, i = 0, \dots, C\}, & j = 0 \\ \{Rank(\{i|\alpha_i = 1, i = 0, \dots, C\})_j\}, & j = 1, \dots, C' \end{cases} \tag{11}$$

where the subscript (i, j) represents the channel. C is the original channel number and C' is the channel number after label or prediction recalibration. Then the marginal Dice loss and CE loss can be calculated as follows:

$$L_{mDice} = -\frac{2p'g'}{p' + g'} \tag{12}$$

$$L_{mCE} = -g' \log(p') \tag{13}$$

Compared with training single-organ networks for each specific organ or training one multi-organ network by partially cross entropy loss (Tang et al., 2018), which are common methods in partially supervised multi-organ segmentation, our method can not only effectively utilize mutual exclusion relationship between different organs, but also work when the annotated classes in each dataset is a proper subset of all annotated classes.

4. Experiments

4.1. Materials

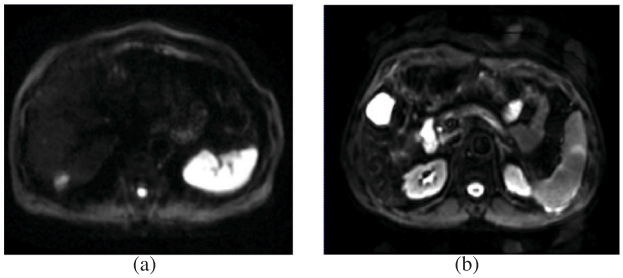
340 CT scans and 200 MRI cases are included in our study. The former are collected from several public dataset: BTCV[1] (47), TCIA[2] (43), Nonalcoholic Fatty Liver Disease[3] (NAFLD, 64), 3D-IRCADb-01[4] (15), Decalthon[5] (69) and Whole abdominal ORgan Dataset[6] (Luo et al., 2022) (WORD, 102). The size of images in these CT datasets is $512 \times 512 \times N$, where N ranges from 26 to 461. The horizontal spacing ranges from 0.5 mm to 1 mm, and the thickness ranges from 0.5 mm to 8 mm. Noted that a few additional organs are manually annotated in some databases (e.g. TCIA, NAFLD) in our experiment, so the labeled and unlabeled organs are not identical to the public database, but majority of them are the same. More details of these 6 public CT datasets and the 11 target organs to segment in this paper are shown in Table 1. MR images are collected from Beijing Friendship Hospital, and Captain Medical University, which are all obtained by 3T MR scanners, including Siemens Prisma, GE 750 W and Philips Ingenia. Each patient underwent dynamic contrast-enhanced MRI (DCE-MRI) and other conventional magnetic resonance imaging at the same examination for the evaluation of the upper abdomen. The DCE-MRI examination was performed using the 3D gradient echo sequence with liver acquisition with volume acceleration protocol within a breath-hold and the following parameters: (1) Siemens Prisma: repetition time/echo time, 3.76 msec/1.23 msec; flip angle, 12°; matrix size, 288 × 61; section thickness/interslice gap, 3 mm/20 mm; field of view (FOV), 325 mm × 100 mm; (2) GE 750 W: repetition time/echo time, 4.1 msec/1.9 msec; flip angle, 12°;matrix size, 288 × 170; section thickness/interslice gap, 4 mm/0 mm; field of view (FOV), 380 mm × 80 mm. (3) Philips Ingenia: repetition time/echo time, 3.6 msec/1.32 msec; flip angle, 15°; matrix size, 216 × 188; section thickness/interslice gap, 4 mm/−2 mm; field of view (FOV), 400 mm × 349 mm. In our study, three DCE-MR sequences, including EAP, PVP and DP, and six conventional MR sequences, including T1 W, IP, OP, T2 W, DWI and apparent diffusion coefficient map (ADC), are enrolled. However, due to the loss and damage of files during data transmission, not all patients are equipped with 9 sequences. The final number of samples included in each sequence is shown in Table 2. Different sequences in our dataset vary a lot in the spacing, especially in the inter-layer spacing. For instance, the inter-layer spacing of T1 W and contrast-enhanced dynamic sequences ranges from 2 mm to 5 mm, and that of IP and OP ranges from 2 mm to 9 mm. However, in regard to T2 W, DWI and ADC, the inter-layer spacing ranges from 6 mm to 10.5 mm.

Table 2
Details of data partition.

	CT	MRI								
		T1 W	EAP	PVP	DP	T2 W	IP	OP	DWI	ADC
Train	203	79	119	117	120	120	116	116	101	100
Validation	69	31	40	40	39	40	39	39	30	30
Test	68	27	39	39	40	40	39	38	32	32

Due to the collected MRI scans focusing on the upper-abdomen, we regard spleen, right kidney (R-kidney), left kidney (L-kidney), gallbladder, liver, stomach, aorta, inferior vena cava (IVC), duodenum,

[1] https://www.synapse.org/#!Synapse:syn3193805/
[2] https://wiki.cancerimagingarchive.net/display/Public/Pancreas-CT/
[3] https://repository.niddk.nih.gov/studies/nafld_adult/
[4] https://www.ircad.fr/research/data-sets/liver-segmentation-3d-ircadb-01/
[5] https://paperswithcode.com/dataset/medical-segmentation-decathlon
[6] https://github.com/HiLab-git/WORD



(a) (b)

Fig. 4. DWIs with poor image quality. (a) DWI with low SNR. (b) DWI with low intensity caused by the severe fatty liver.

Table 3
Architecture of the refactored generator.

Layer	Stride	Norm	Repeat	Output shape
Image x	–	–	–	256 × 256 × 1
Conv7 × 7	1	IN	1	256 × 256 × 64
Conv4 × 4	2	IN	1	128 × 128 × 128
Conv4 × 4	2	IN	1	64 × 64 × 256
ResBlock	1	IN	4	64 × 64 × 256
ResBlock	1	AdaIN	4	64 × 64 × 256
UpSample	2	–	1	128 × 128 × 256
Conv3 × 3	1	AdaIN	1	128 × 128 × 128
UpSample	2	–	1	256 × 256 × 256
Conv3 × 3	1	AdaIN	1	256 × 256 × 64
Conv7 × 7(1)	1	–	1	256 × 256 × 1

pancreas and esophagus as our target organs. CT and MRI cases are all randomly divided into train set, validation set and test set by the ratio of 3:1:1. More details of the data partition is displayed in Table 2. The organ masks of the scans in the test set are annotated by two radiologists. All target organs are delineated in the T1 W and dynamic-contrasted enhanced sequences (EAP, PVP, DP) while only liver mask is delineated in T1 W, IP, OP and DWI. Note that ADC is calculated from DWIs with different b values, which means ADC share the same organ mask with its corresponding DWIs, thus there is no need to segment ADC anymore. Consequently, in this paper, we present the generation results for all 9 sequences, and show the segmentation results for the remaining 8 sequences except ADC. Low intensity caused by the presence of severe fatty liver or low signal-to-noise ratio (SNR) caused by the high B-value will lead to the blurred liver boundaries in DWIs (4), which affects the annotation of liver. In the test set, five DWIs have poor image quality due to the above reasons and are not included in the segmentation result statistics.

The following preprocessing operations are implemented to unify the CT and MRI images in our experiments. As mentioned above, the MR scans included in our study focus on the upper abdomen, while CT volumes in some datasets, e.g. WORD, contain the full abdomen. To ensure that CT images and MR images have similar regions of interest (ROIs), the upper and lower boundary of the organ masks (with an expansion of 5 slices in both directions) are used to determine the vertical boundary of upper-abdomen for CT volumes. After that, the CT volume is binarized by −175 Hounsfield Unit (HU), and the maximum connected region is detected to determine the location of the body. According to the vertical boundary and the maximum connected region, the final ROI of the CT volume can be obtained. In terms of MR volumes, there is no need to detect the vertical boundary. Similar binary thresholding and morphological operations are implemented to locate the ROI, with the threshold set to 100 if the 99.9 percentile of the volume intensities is greater than 400, and 40 otherwise. Second, normalization is implemented on both 3D CT volumes and MR volumes. For CT volumes, the intensity is first truncated at [−175, 350] Hounsfield Unit (HU), and then rescaled to $[-1, 1]$. MR volumes are normalized similarly with truncation threshold of [0.05, 99.5] percentiles. Finally, both CT images and MR images are resized to 256×256.

4.2. Implementation details

1. Network Architecture
   The architecture of the refactored generator is shown in Table 3. The encoder and decoder of this generator are roughly symmetrical and both of them are composed of 3 convolution blocks and 4 residual blocks. To reduce the loss of spatial information and ensure the field of reception, 2 down-sampling operations are included in this architecture. The architectures of style encoder, mapping network and discriminator are implemented following StartGAN v2 and we modified the number of out branches as 10, one for CT and 9 for the MR sequences.
2. Hyperparameters
   In the OMDG framework, the number of output branches of the style encoder, mapping network and discriminator is 10, which equals to the number of sequences (CT, T1 W, EAP, PVP, DP, T2 W, IP, OP, DWI, ADC) used in our experiment. Adam optimizer is utilized to optimize the network where the learning rate is initialized to 1e−6 for the mapping network and 1e−4 for other networks. The batch size is set to 8 and the maximum iteration is set to 80000 in this study. Other hyper-parameters are set the same as those in StarGAN v2.
   In the segmentation stage, we use 3D nnUNet as the network structure due to its remarkable performance on different datasets (Isensee et al., 2019a). Besides that the batch size is set to 8 and the nnUNet built-in postprocessing is discarded, the other settings are kept default in the self-configuration framework of nnUNet. To improve the training efficiency, we divide the 8 MR sequences (as mentioned in Section 4.1, ADC shares the organ mask with its corresponding DWI, thus no additional segmentation of ADC is required) into three groups based on the similarity of imaging mechanisms and thickness ranges, and train a segmentation network for each group.
   - Group 1: T1 W, EAP, PVP, DP
   - Group 2: IP, OP
   - Group 3: T2, DWI
3. Resolution Unification
   The synthetic MR images share the same spacing with their corresponding source CT images (256×256). However, the spacing between CT images and MR images varies greatly. If these synthesized MR images are used directly to train nnUNet, the optimal image spacing planned by nnUNet is based on the spacings of CT images, which may damage the segmentation performance on real MR images. Therefore, resolution unification is employed to keep the resolutions of the synthetic MR images and the real MR images similar. First, the synthetic MR images (256×256) are resampled to the original spacing of their corresponding CT images, i.e., the spacing before image resizing. Then according to the distribution of the spacing of each sequence in our MR dataset, the synthetic MR images are further resampled, thus these resampled synthetic MR images share a similar image resolution with the real MR images. According to Fig. 5, there is no significant difference between the distributions of the spacing in the resampled synthetic MR images and real MR images (Mann–Whitney U test, $p > 0.05$).
4. Postprocessing
   Morphological operations are conducted to fill holes in the spleen and liver. In addition, postprocessing based on connected components is implemented to keep the largest connected region or remove small isolated false positives for segmented organs.

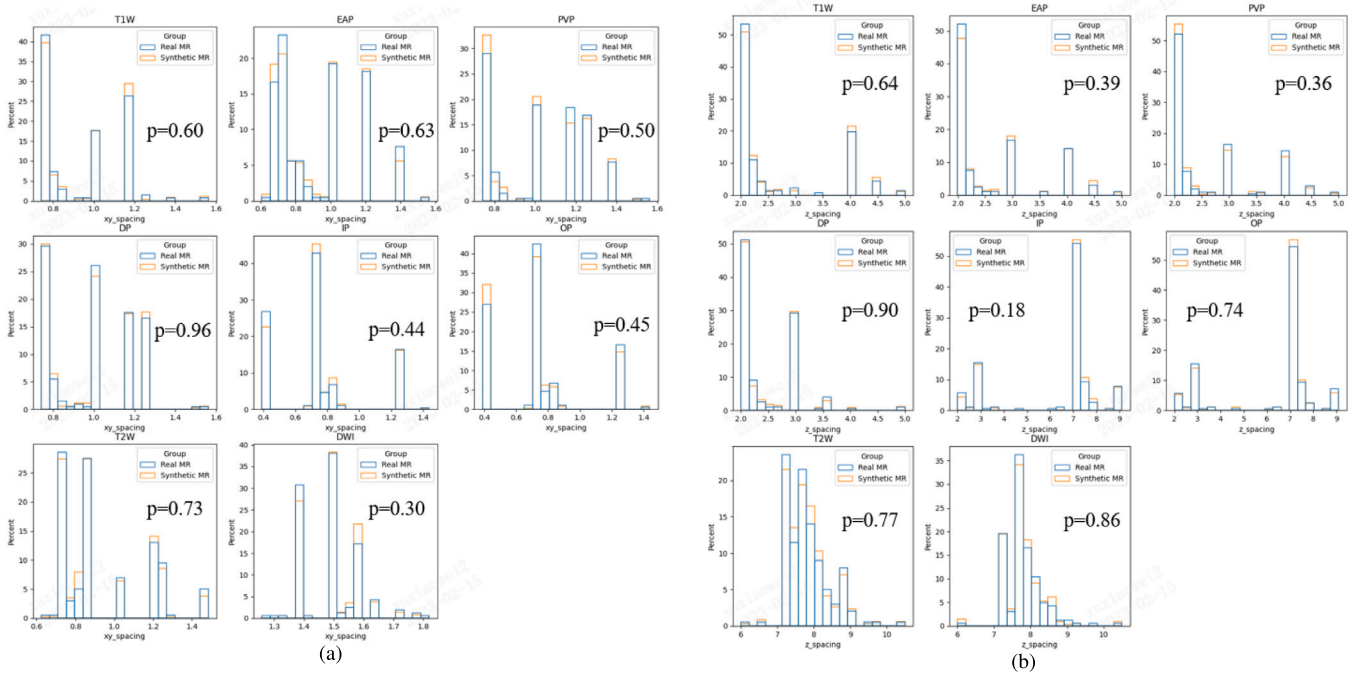


Fig. 5. The spacings of different sequences in resampled synthetic MR images and real MR images. (a) The spacings of the X-axis or the Y-axis in different sequences. (b) The spacings of the Z-axis in different sequences.

4.3. Results

The CT-to-multi-sequence MR image translation results are qualitatively displayed in Fig. 6, where Fig. 6(a) shows the generated images whose styles are extracted from the reference images while Fig. 6(b) shows the generated images whose styles are the mapping results of latent vectors. In terms of visual perception, the generated images actually learn some domain-specific features. For instance, the aorta in the generated EAP images is highlighted, and the edge effect in the pseudo OP images is obvious.

DSC and 95% Hausdorff Distance (HD_{95}) are utilized to evaluate the segmentation performance. DSC, a volume-based evaluation metric, is sensitive to the internal segmentation of the target organs and is formulated as follow:

$$DSC(S_g, S_p) = \frac{2 \times |S_g \cap S_p|}{|S_g| + |S_p|} \tag{14}$$

where S_g and S_p are the ground-truth and predicted mask respectively. HD_{95}, a variant of HD, calculates the 95th percentile of the distances between boundary points of S_g and S_p to eliminates the impact of a very small subset of the outliers. It is calculated as:

$$HD_{95}(G, P) = \max\{\max_{g \in G}{}_{95\text{th}} \min_{p \in P} d(g, p), \max_{p \in P}{}_{95\text{th}} \min_{g \in G} d(g, p)\} \tag{15}$$

where G and P are the boundary point sets of S_g and S_p respectively. According to the equation, HD_{95} is invalid when over-segmentation occurs in these completely resected organs, thus organs completely resected by surgery were not included in the quantitative analysis. In our dataset, three patients underwent left nephrectomy, right nephrectomy and cholecystectomy respectively. The quantitative segmentation results of our methods are shown in Table 4. According to it, the segmentation performance of the network on left kidney is worse than that on right kidney. This phenomenon is mainly caused by an extreme case where the left kidney is severe atrophied and the prediction is a heavy under-segmentation. Moreover, due to the contrast agent, the aorta in EAP, PVP and DP is highlighted and has better contrast with surrounding tissue, which leads to higher DSCs of aorta in those sequences than T1 W scanned before contrast agent injection. Similarly, the contrast agent flows into the IVC since PVP, therefore, the segmentation performance of IVC is better in PVP and DP.

4.4. Ablation study

4.4.1. Evaluation of the refactored generator

Fig. 7 illustrates the source CT images and the generated multi-sequence MR images of different generator architectures. For facilitating visual comparison of different methods in organ geometry preservation, organ profiles are superimposed on the CT and MR images. According to the generation results in the second and third columns, compared with StarGAN v2, the refactored generator (OMDG w/o SC) shows apparent advantages in maintaining the geometry of target organs. For instance, the geometry of kidneys (amplified in the white box) in the DP image (second row) generated by StarGAN v2 is quite different from that in the source CT image.

Due to the lack of paired CT and multi-sequence MR images, it is difficult to quantitatively evaluate the geometry-consistency. Empirically, if the anatomical structure is destroyed during image translation, the organ mask of the source CT image will not match the generated MR images, thus impairing the segmentation performance. Therefore, once other factors are controlled (e.g., segmentation network, hyperparameters), the segmentation results in the subsequent stage can verify the effectiveness of the new generator architecture to some extent. According to Table 5, the DSC and HD_{95} values after generator refactoring are significantly better than that before refactoring, which may reflect better geometry-consistency of this new generator architecture.

Furthermore, we assess the anatomical structural consistency based on image registration. We assume that the smaller the Jacobian determinant of the deformation field between the generated MR image and the source CT image, the higher the geometric consistency between them. Based on this hypothesis, deformable registration based on ANTS (Avants et al., 2009)[7] is employed between generated MR

[7] https://github.com/ANTsX/ANTsPy/

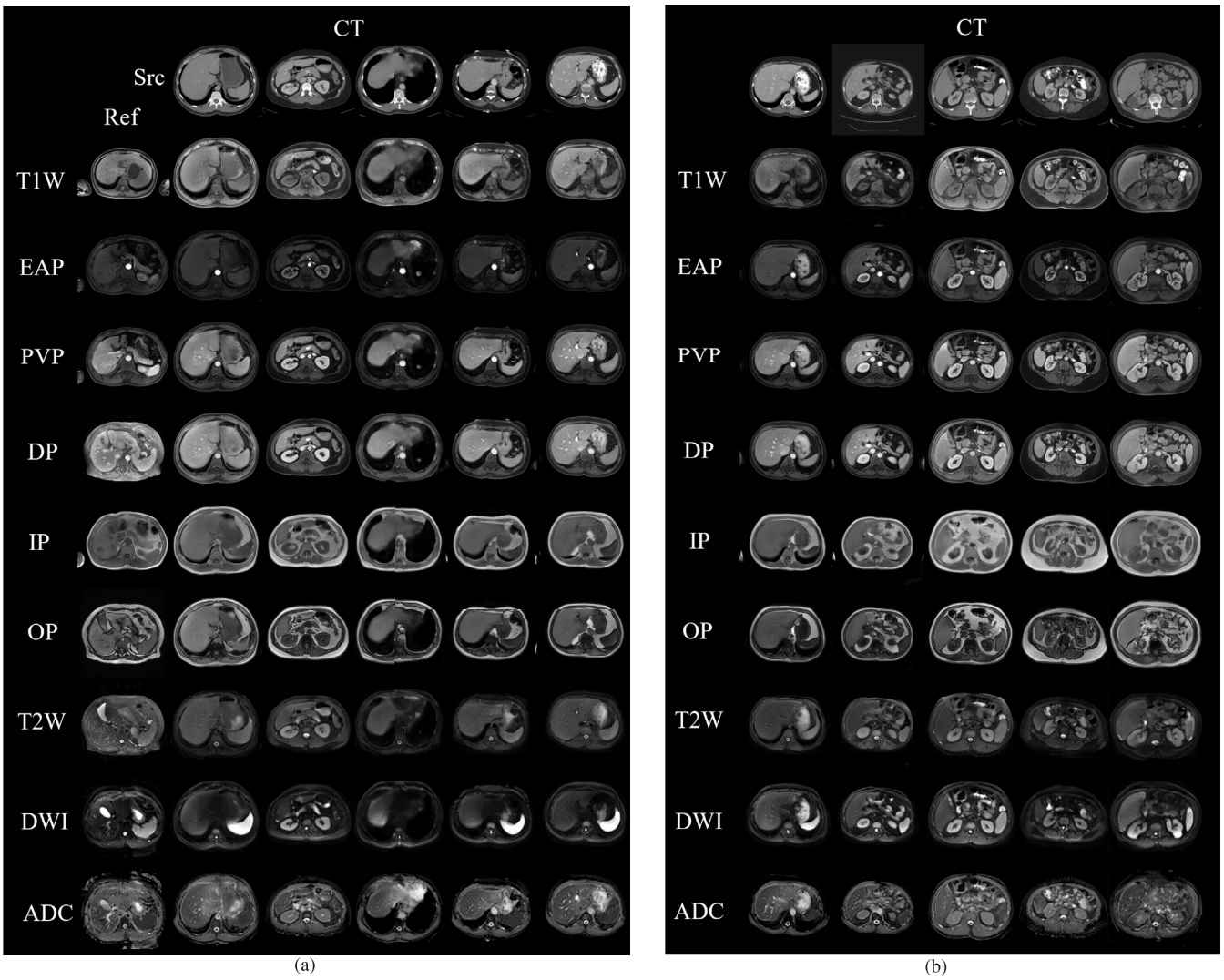


Fig. 6. CT-to-Multi-Sequence MR image translation result. (a) Translation results with style codes extracted from the reference images. The first row and column display the source images and reference images respectively. And the rest images are generated images which theoretically have the same geometry as the source image in the same column and the same style as the reference image in the same row. (b) Translation results with style codes mapped from latent vectors. The first row shows the source images.

Table 4
Quantitative segmentation results of OMUDA in terms of DSC (%) and HD_{95} (mm)

		Spleen	R-Kidney	L-Kidney	Gallbladder	Liver	Stomach	Aorta	IVC	Duodenum	Pancreas	Esophagus	Avg.
DSC	T1 W	94.12 ±2.09	94.92 ±1.15	91.12 ±16.82	80.85 ±11.50	96.43 ±0.73	87.50 ±6.60	85.95 ±4.89	68.29 ±13.20	76.51 ±7.35	81.35 ±7.43	65.49 ±18.80	83.81 ±14.24
	EAP	91.38 ±7.19	95.39 ±0.92	92.64 ±11.15	76.31 ±20.96	95.42 ±1.03	85.44 ±8.36	91.05 ±2.73	67.23 ±14.14	74.47 ±11.84	83.41 ±9.27	62.54 ±17.85	83.17 ±15.85
	PVP	94.97 ±1.86	95.54 ±0.87	92.33 ±15.22	73.33 ±25.88	96.41 ±1.03	86.13 ±7.85	92.51 ±2.15	82.35 ±7.25	75.52 ±10.92	83.29 ±10.16	74.53 ±12.55	86.08 ±14.06
	DP	95.11 ±1.80	94.73 ±0.85	93.26 ±9.09	73.90 ±21.76	96.52 ±0.79	87.59 ±8.01	92.54 ±2.28	81.79 ±6.13	76.46 ±9.91	81.47 ±11.28	80.50 ±5.87	86.73 ±12.01
HD_{95}	T1 W	3.53 ±2.93	2.45 ±0.77	3.80 ±5.34	5.80 ±3.39	3.21 ±1.14	10.31 ±11.69	4.60 ±3.39	9.69 ±7.19	18.90 ±20.54	6.11 ±5.35	13.42 ±14.15	7.47 ±10.33
	EAP	5.11 ±5.43	2.31 ±0.61	3.03 ±2.60	9.19 ±10.02	3.98 ±1.10	12.08 ±15.06	2.34 ±1.15	10.72 ±8.35	16.42 ±17.20	7.11 ±10.06	13.51 ±20.79	7.82 ±11.73
	PVP	2.83 ±2.10	2.06 ±0.40	3.45 ±5.80	11.28 ±15.61	3.29 ±2.05	13.47 ±17.85	2.15 ±1.30	5.55 ±5.48	15.87 ±17.92	6.35 ±8.90	5.95 ±4.88	6.58 ±10.84
	DP	2.72 ±2.11	2.65 ±0.72	3.02 ±3.20	9.31 ±9.23	3.13 ±0.88	11.07 ±17.32	2.55 ±3.01	5.43 ±5.36	11.85 ±14.78	7.23 ±9.11	4.27 ±2.15	5.76 ±8.92

		IP	OP	T2 W	DWI
Liver	DSC	91.45 ± 2.15	94.35 ± 1.79	87.81 ± 13.41	87.09 ± 8.45
	HD_{95}	7.83 ± 4.09	5.25 ± 2.90	16.40 ± 15.76	14.62 ± 11.73

mean ± *std*.

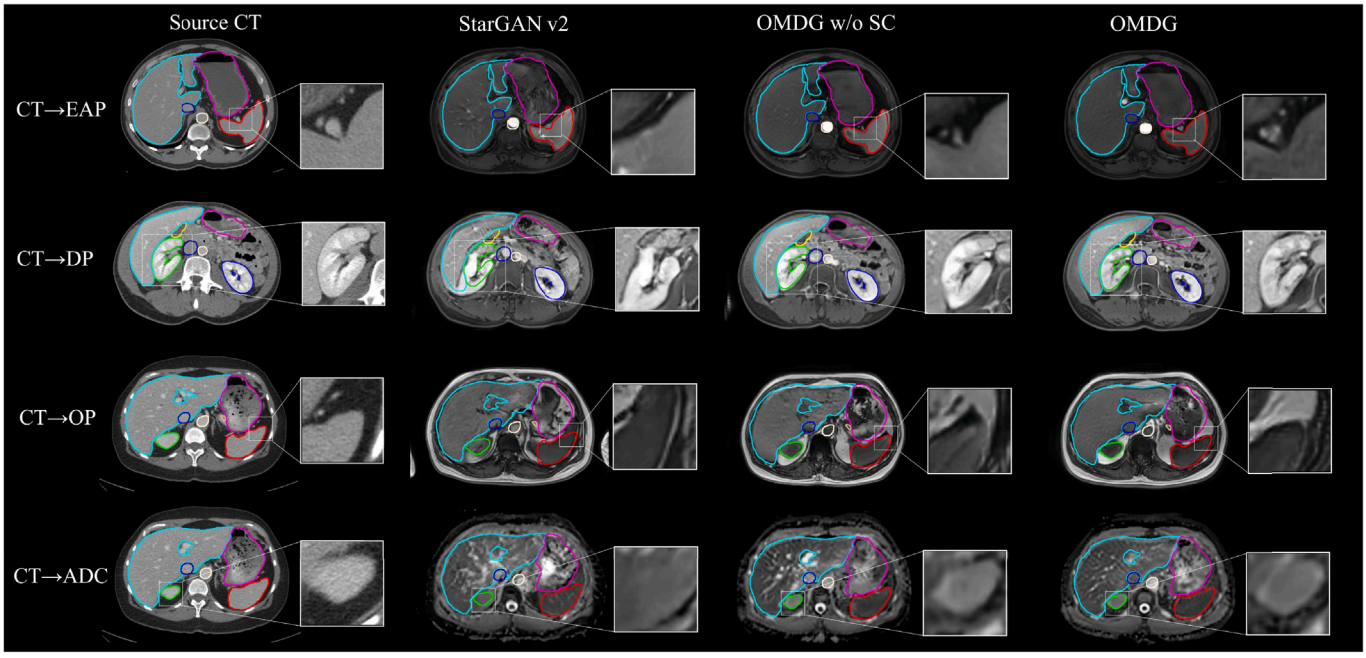


Fig. 7. Qualitative evaluation of different generators. The first row is the source CT images. The second column displays the generation results of StarGAN v2, and the third column (OMDG w/o SC) is the generation results when the original generator in StarGAN v2 is replaced by the refactored architecture. The last column (OMDG) shows the generated multi-sequence MR images of our OMDG framework.

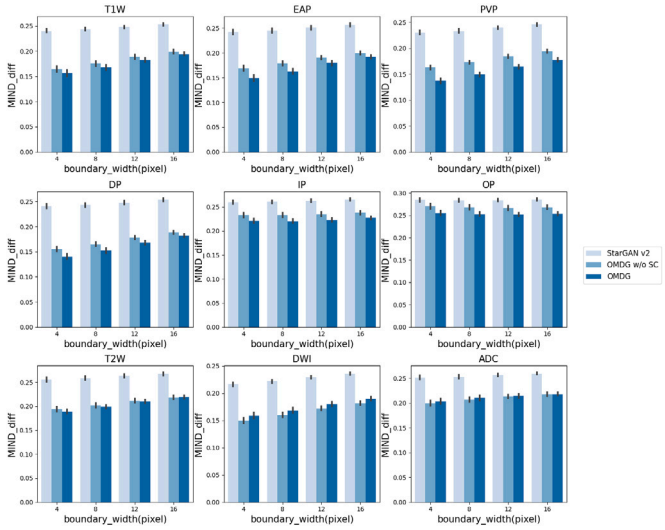


Fig. 8. The average MIND differences across all target organs between source CT images and generated multi-sequence MR images for different methods.

Table 5
Average DSC (%) and HD_{95} (mm) of different sequences in ablation study. Baseline is StarGAN v2, and Ref_G and SC mean the refactored generator and style constraint respectively.

Modules		Metrics	Sequences							
Ref_G	SC		T1 W	EAP	PVP	DP	IP	OP	T2 W	DWI
		DSC	70.11	73.12	75.32	76.68	85.92	88.35	71.59	43.10
✓		DSC	82.61	82.53	85.04	84.94	91.24	94.09	61.74	63.42
✓	✓	DSC	**83.81**	**83.17**	**86.08**	**86.73**	**91.45**	**94.35**	**87.81**	**87.09**
		HD_{95}	17.38	14.07	10.73	10.27	20.03	15.24	34.27	70.52
✓		HD_{95}	8.01	8.42	**6.25**	6.66	**7.15**	5.32	42.77	35.33
✓	✓	HD_{95}	**7.47**	**7.82**	6.58	**5.76**	7.83	**5.25**	**16.40**	**14.62**

Table 6
Structural consistency evaluation based on deformation field for different sequence. Baseline is StarGAN v2, and Ref_G and SC mean the refactored generator and style constraint respectively.

Modules		Sequences									Avg.
Ref_G	SC	T1 W	EAP	PVP	DP	IP	OP	T2 W	DWI	ADC	
		0.166	0.187	0.180	0.179	0.156	0.151	0.175	0.166	0.182	0.172
✓		**0.144**	0.147	0.145	0.140	**0.141**	0.150	0.179	**0.163**	**0.169**	0.153
✓	✓	0.145	**0.141**	**0.138**	**0.139**	0.142	**0.149**	**0.168**	0.164	0.173	**0.151**

volumes and their corresponding source CT volumes to obtain the deformation field, and then the L1-norm of the Jacobian determinant of the deformation field is calculated. Note that affine and deformable transformations with mutual information as optimization metric are used in this experiment. According to the results in Table 6, the average L1-norm of OMDG w/o SC is smaller than that of StarGAN v2 (0.153 vs 0.172), indicating the effectiveness of the refactored generator in geometry-consistency.

Then, we utilize Modality Independent Neighbourhood Descriptor (MIND) (Heinrich et al., 2012) to compare the structural features between source images and generated images. MIND is defined using a non-local patch-based self-similarity and depends on local image structures instead of intensity values, thus this descriptor is insensitive to image intensities and suitable to illustrate the structure feature of different modality images. Owing to the fact that organs may show different textures in different sequences, for instance, the portal vein in liver is highlighted in PVP and DP but not in T1 W and EAP, we focus more on the structural difference on the boundary of target organs. Fig. 8 shows the structure evaluation results based on MIND. From it, we can see that the new generator (OMDG w/o SC) has less structural difference than the original one (StarGAN v2), verifying a better geometry-consistency of it.

4.4.2. Evaluation of the style constraint

Style constraint is imposed on the style codes to explicitly minimize the inter-class variance and maximize the intra-class variance, and thus reduce domain aliasing. In accordance to Fig. 9 that shows the t-SNE maps of the style codes generated by the style encoder before and after

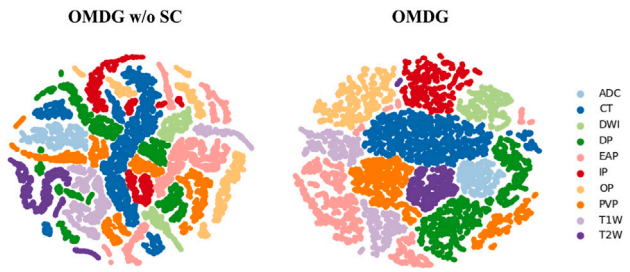


Fig. 9. t-SNE maps of style codes generated by style encoder in OMDG w/o SC and OMDG.

style constraint, although the multiple output branches implicitly aggregate the styles from the same domain to some extent, the imposition of style constraints make the styles from the same domain cluster more tightly, which contributes to disentangle styles from contents. In other words, style constraint can compel the style encoder to extract similar styles from different slices belonging to the same modality, thus more pure styles are extracted and less content information is involved in the style codes. Its contributions to the generation task and subsequent segmentation task are shown in Fig. 8 and Table 5. In Fig. 9, the average MIND difference between the source CT image and the generated MR images are reduced for most sequences (e.g., T1 W, EAP, PVP, DP and IP) after imposing the style constraint. In Table 5, after the imposition of style constraint, the average DSCs for all sequences are improved and the average of $HD_{95}s$ for a majority of the sequences are decreased.

4.5. Comparison study

We compare OMUDA with other state-of-the-art unsupervised synthesis methods: SIFA (Chen et al., 2020), CycleGAN (Zhu et al., 2017), MUNIT (Huang et al., 2018) and StarGAN v2 (Choi et al., 2020). Except SIFA, all other methods can only complete cross-domain image translation, thus we train their exclusive nnUNets from scratch based on MR images synthesized by these methods. Owing to the one-to-one image translation property of SIFA, CycleGAN and MUNIT, 8 models are supposed to be trained for CT-to-multi-sequence MR image translation in these methods.

1. SIFA: It is an unsupervised domain adaptation framework which can adapt a segmentation network to an unlabeled target domain by conducting synergistic alignment of domains from both image and feature perspectives. SIFA integrates segmentation with image synthesis, thus no additional segmentation networks need to be trained.
2. CycleGAN: It is a classic unpaired image-to-image translation framework which for the first time employs cycle-consistent adversarial networks to handle the image translation issue in the absence of aligned image pairs.
3. MUNIT: Different from CycleGAN, MUNIT explicitly decomposes the image representation into a content code that is domain-invariant, and a style code that captures domain-specific properties. Therefore, domain translation can be achieved by replacing the style code of the source image by the style codes of the target domain.
4. StarGAN v2: It is a diverse image synthesis framework for multiple domains, which is the baseline of OMDG. In this comparison study, all configurations of StarGAN v2 remain in the original state.

The implementation of these methods are all borrowed from the github code database,[8][9][10] expect SIFA[11] which is re-implemented in PyTorch by ourselves based on the TensorFlow version.[12] More details about the settings and hyper-parameters of these comparative experiments can be found in the Appendix.

4.5.1. Segmentation performance on internal testing set

The DSCs and HD_{95}s for different sequences of these methods are shown in Fig. 10. Comprehensive consideration of DSC and HD_{95}, OMUDA outperforms SIFA, MUNIT and StarGAN v2 by a large margin for multiple organs in multiple MRI sequences ($p \leq 0.05$, T-test for normally distributed data, Wilcoxon signed-rank test for otherwise). Although SIFA integrates image synthesis with segmentation and makes the training more efficient, it does not perform well in our task. The main reasons for its poor segmentation performance are the poor quality of image alignment (e.g., the generated PVP image in Fig. 11 is very similar in appearance to the source CT image and retains too many features of the source CT image) and the limited field of view of its 2D segmentation network (cross-axial information cannot be fully utilized). In contrast to SIFA, OMUDA generates images with higher fidelity (last column in Fig. 11) and employs 3D segmentation networks to make full use of the 3D content information by separating the segmentation from the image synthesis. MUNIT and StarGAN v2 make remarkable achievements in natural image synthesis, however, they focus more on the fidelity or diversity of the generated images but less on cross-domain structural consistency which is critical for the subsequent medical image segmentation task (the fourth and fifth columns in Fig. 11). In this regard, OMDG based on StarGAN v2 puts more emphasis on the structural consistency during cross-domain image translation from the aspects of the generator architecture and extra constraint, which significantly improves the segmentation performance (Fig. 10). According to the quantitative results, the segmentation performance of OMUDA is inferior to that of CycleGAN in a majority of the organs in T1 W and dynamic-contrast enhanced MR sequences, but is significantly superior ($p \leq 0.05$) to that of CycleGAN in modalities which have even more different appearance from the source CT images, e.g., IP, T2 W, DWI. This reflects that our method is more advantageous in image translation problems with large inter-domain differences.

It should be noted that though the segmentation performance of OMUDA is not the best for all organs in all sequences, it has apparent advantages in training efficiency which will be detailedly illustrated in Section 4.5.3. In other words, the highlight of OMUDA is its comprehensive consideration of segmentation and training efficiency.

4.5.2. Segmentation performance on external testing set

Besides evaluating the segmentation performance on the internal testing set, we further assess the robustness of the segmentation networks trained by different methods on two external datasets: Multi-Modality Abdominal Multi-Organ Segmentation Challenge 2022 (AMOS22)[13] and Combined Healthy Abdominal Organ Segmentation (CHAOS).[14] 40 MR images [(39 T1 W or DCE-MR images and 1 OP image)] in the train set of AMOS22 are used as our external testing set and the segmentation results of different methods on them are displayed in Table 7. It should be noted that due to the property of HD_{95}, only the organs which have both ground-truth (i.e., organs which are not resected) and predictions (i.e., organs predicted by all methods) are used to calculate the quantitative results. Comprehensively considering

[8] https://github.com/harveerar/PMB_2020_Self_Derived_OAD_Segmentor/tree/master/PMB_attention_code
[9] https://github.com/NVlabs/MUNIT
[10] https://github.com/clovaai/stargan-v2
[11] https://github.com/JianghaoWu/SIFA-pytorch.git
[12] https://github.com/cchen-cc/SIFA
[13] https://amos22.grand-challenge.org/
[14] https://chaos.grand-challenge.org/

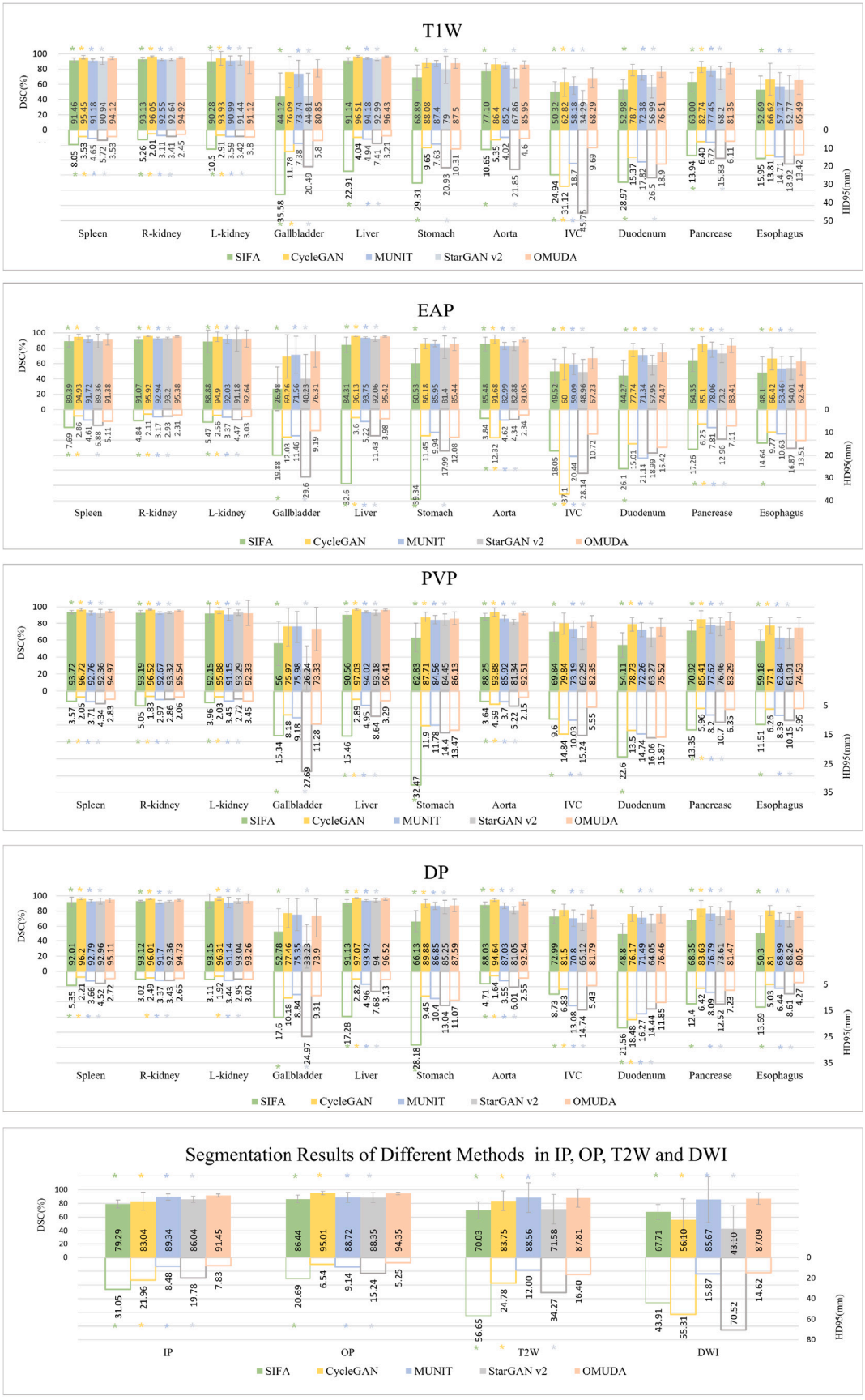


Fig. 10. Segmentation results of different methods on the internal testing set. * represents that the results achieved by this method are significantly different ($p \leq 0.5$) from those achieved by OMUDA.

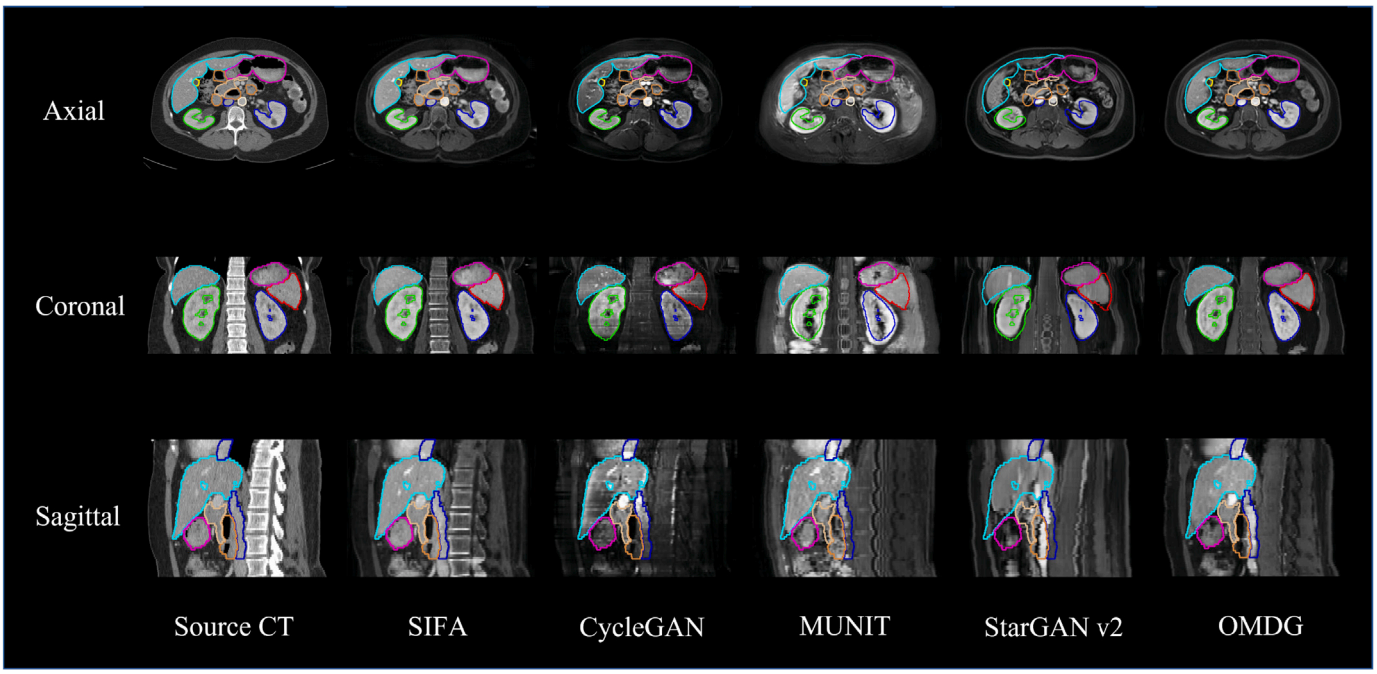


Fig. 11. Generated PVP images of different methods. The first column is the source CT image shown in axial, coronal and sagittal view. The following columns are the pseudo PVP images generated by SIFA, CycleGAN, MUNIT, StarGAN v2 and OMDG respectively.

of DSC and HD_{95}, although the segmentation of CycleGAN is slightly superior in a majority of the organs, OMUDA also achieves satisfactory results on the AMOS22 dataset.

Different from AMOS22 which mainly focuses on evaluating the segmentation performance on DCE-MR images, CHAOS challenge aims at the segmentation of abdominal organs in IP, OP and T2 W. In this challenge, the test set contains 20 subjects, and Dice, Average Symmetric Surface Distance (ASSD), Maximum Symmetric Surface Distance (MSSD) and Relative Absolute Volume Difference (RAVD) are adopted to evaluate the segmentation performance on these scans. Table 8 shows the segmentation results achieved by different methods in the CHAOS challenge, where the best result of Task 3 on the leaderboard is also displayed. First, among the comparison algorithms involved in our experiment, OMUDA ranks first in OP and T2 W, and ranks second in IP, indicating its satisfactory segmentation ability in these MR modalities. Second, compared with the state-of-the-art (SOTA) fully supervised segmentation method, there is still a gap in the segmentation performance between OMUDA and it. However, OMUDA has not been trained with a single real MR image, and the segmentation network is only based on the benchmark nnUNet. Contrarily, the top player on the leaderboard not only train the segmentation network with real MR images in a fully supervised manner, but also use some ensemble tricks, e.g., integrating the output of different models and integrating the segmentation results of IP and OP. Thus the performance gap can be reduced when the segmentation network in OMUDA is fine-tuned by real MR images or more segmentation tricks are employed in the training or inference phase.

According to the quantitative results on these two external test sets, OMUDA and CycleGAN have their own advantages in different MR modalities. On the basis of ensuring segmentation performance, our OMUDA has an obvious advantage in the training efficiency, as detailed in the following.

4.5.3. Generation performance

Due to the dilemma of lacking paired CT-MR data, it is hard to quantitatively evaluate the generation results of different methods. Settle for the second best, we compare the intensity distribution of the synthetic MRIs and the real MRIs for each sequences. Correlation coefficient is used to measure the similarity of two histograms. According to Table 9, OMUDA achieves the highest correlation coefficient among these methods. The synthetic MR images generated by OMUDA share more similar intensity distribution with the real MR images in the most sequences, that is T1 W, EAP, PVP, DP, IP, OP.

4.5.4. Training efficiency

To better analyze the efficiency of our proposed framework, the model parameters and FLOPs (float point operations) of OMDG and CycleGAN are shown in Table 10, where the FLOPs of CycleGAN is the accumulated FLOPs in 9 image translation tasks, i.e., CT-T1 W, CT-EAP, CT-PVP, CT-DP, CT-T2 W, CT-IP, CT-OP, CT-DWI and CT-ADC. According to the model parameters, for the sake of translating CT to 9 MR sequences, CycleGAN are supposed to optimize 7.69 times as many parameters as OMDG. In terms of FLOPs, OMDG has obvious advantages in both training and inference stages, reducing floating-points calculations by around 87 and 30 percents, respectively.

In order to further illustrate the efficiency of OMDG in training, we display the generated images during iterations (Fig. 12) to visually compare its speed of convergence with CycleGAN. It can be seen that OMDG and CycleGAN all fail to generate fidelity MR images at the beginning of the iteration (iteration $\times$ batch size = 32000), e.g., the intensity of the rib or vertebra is not suppressed (highlighted by red arrows) and the gridding artifacts exist in generated images (highlighted by the green box). In comparison, OMDG converges slightly faster than CycleGAN. For instance, the liver in T1 W and OP generated by CycleGAN at the beginning of the iteration (highlighted by the yellow box) is very similar as that in the source CT images, but this is not the case for OMDG. Moreover, within the iterations we visualized, OMDG is supposed to generate IP images with a similar tissue-contrast to real IP images faster than CycleGAN (highlighted by the red box), which is consistent with the quantitative result in Table 9. All of those may demonstrate that OMDG has some advantages in convergence speed. At least compared with CycleGAN, there is no obvious disadvantages. To sum up, from the aspects of FLOPs and the speed of convergence, OMDG is more efficient in training and inference than CycleGAN. Lastly, it should be noted that when more domains are included in mutual image translation, the efficiency of OMDG in both training and inference stage will be more prominent.

Table 7
Quantitative segmentation results of different methods on AMOS22 in terms of DSC (%) and HD_{95} (mm).

		Spleen	R-Kidney	L-Kidney	Gallbladder	Liver	Stomach	Aorta	IVC	Duodenum	Pancreas	Esophagus
DSC (%)	SIFA	86.94 ±11.72[a]	88.89 ±11.50[a]	89.01 ±10.18[a]	48.10 ±30.60[a]	90.91 ±2.67[a]	57.91 ±19.55[a]	78.19 ±15.54[a]	60.28 ±15.85[a]	42.15 ±12.59[a]	68.06 ±12.44[a]	41.52 ±21.11[a]
	CycleGAN	**95.91 ±1.28**[a]	**94.24 ±5.63**[a]	**94.55 ±1.92**[a]	79.25 ±17.58	**96.42 ±0.67**[a]	**85.40 ±12.66**[a]	**90.89 ±4.74**[a]	74.08 ±13.51	**63.06 ±11.87**[a]	**82.03 ±8.22**[a]	60.45 ±14.31
	MUNIT	91.89 ±2.56[a]	90.79 ±5.85[a]	90.62 ±2.64[a]	78.80 ±13.61[a]	93.48 ±1.20[a]	82.48 ±10.21[a]	83.98 ±3.37[a]	70.60 ±7.83[a]	60.67 ±11.29	76.51 ±8.88[a]	54.02 ±8.82[a]
	StarGAN v2	92.06 ±3.11[a]	91.09 ±3.91[a]	91.27 ±3.06[a]	62.92 ±21.67[a]	94.53 ±1.17[a]	81.76 ±15.34[a]	71.94 ±10.03[a]	58.61 ±11.20[a]	47.65 ±13.29[a]	73.88 ±8.95[a]	51.36 ±13.50[a]
	OMUDA	95.08 ±1.52	92.50 ±8.63	93.56 ±2.27	**81.46 ±13.00**	96.01 ±0.74	83.13 ±18.59	89.41 ±4.63	**76.06 ±9.59**	59.79 ±12.22	79.16 ±8.91	**62.55 ±12.96**
HD_{95} (mm)	SIFA	12.85 ±14.38[a]	6.69 ±8.69[a]	8.35 ±7.99[a]	20.99 ±16.10[a]	20.21 ±14.26[a]	36.56 ±27.97[a]	18.15 ±25.69[a]	22.81 ±23.31[a]	24.04 ±11.24[a]	12.96 ±9.14[a]	20.53 ±16.27[a]
	CycleGAN	**2.15 ±0.88**[a]	**3.56 ±3.47**[a]	**3.80 ±1.96**[a]	6.76 ±5.85	3.69 ±1.57	11.34 ±19.61	4.94 ±8.91[a]	17.95 ±23.19	**13.54 ±8.08**[a]	**5.85 ±5.65**	8.53 ±6.00
	MUNIT	3.85 ±1.69[a]	4.46 ±2.91[a]	4.50 ±1.47	6.45 ±7.19	5.45 ±2.32[a]	**9.81 ±11.37**	**4.10 ±2.29**[a]	**9.24 ±9.56**	14.86 ±6.99	7.20 ±6.01[a]	**7.51 ±2.99**
	StarGAN v2	3.74 ±1.41[a]	5.13 ±3.11[a]	5.62 ±2.37[a]	13.86 ±11.69[a]	4.69 ±1.43[a]	15.14 ±21.02[a]	16.72 ±15.83[a]	22.41 ±18.87[a]	22.23 ±14.29[a]	9.49 ±8.92[a]	12.12 ±8.70[a]
	OMUDA	2.50 ±0.73	4.76 ±7.40	4.63 ±2.52	**5.68 ±5.02**	**3.61 ±0.97**	10.38 ±16.72	4.52 ±6.20	11.85 ±13.71	15.29 ±7.79	6.41 ±5.61	10.86 ±23.46

mean ± *std*.
[a]Represents that the results achieved by this method are significantly different ($p \leq 0.05$) from those achieved by OMUDA.

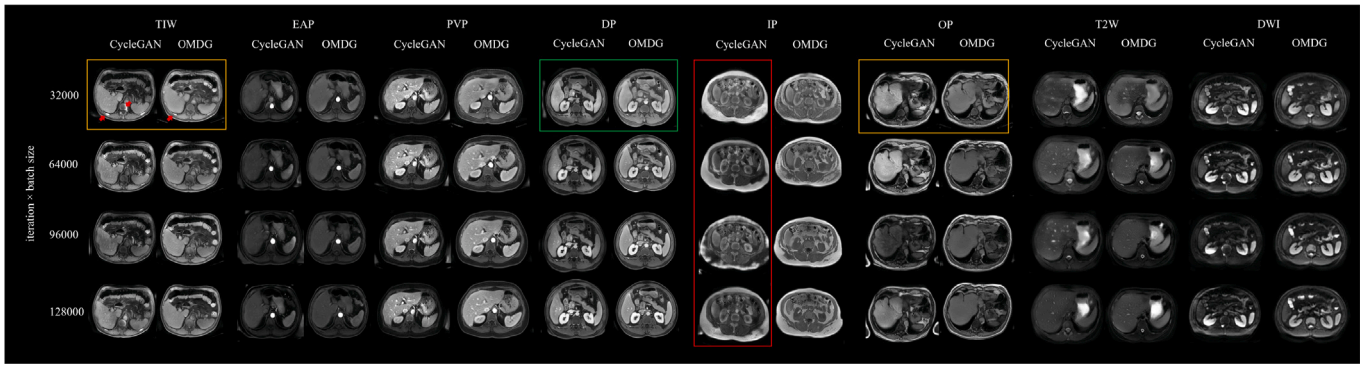


Fig. 12. Visualization of the generated images of CycleGAN and OMDG during iteration. The intensity of the rib or vertebra highlighted by red arrows in the yellow box is not suppressed. Images highlighted by the green box has grid artifacts. In yellow box, the livers in T1 W and OP images generated by CycleGAN are similar with those in source CT images.

4.6. Annotation consistency

To illustrate the reliability of the annotations, inter-observer variability between the annotators is investigated. We randomly selected 10 cases from the test set and had two radiologists re-label the target organs. Each radiologist was assigned five cases that were annotated by the other radiologist before. Fig. 13 shows the inter-observer consistency in different organs, where substantive organs, e.g., liver, spleen and kidney, show high annotation consistency between annotators. Comparatively, some tubular organs, such as IVC and esophagus, have relatively large inter-observer variance, possibly due to their poor contrast in some sequences, such as T1 W. Overall, according to the quantitative results, the annotation standard of these two radiologists is relatively uniform.

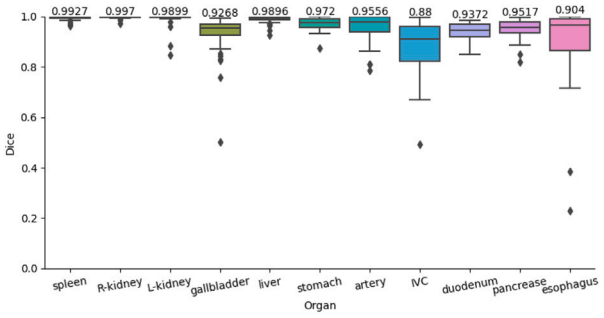


Fig. 13. Inter-observer consistency of organ annotation in our experiment.

5. Discussion

In this study, an image-synthesis-based and annotation-free segmentation framework is proposed to segment the abdominal organs in multiple MR sequences with no annotation provided in these sequences. This proposed OMUDA contains a one-to-multiple domain adaption generation framework (OMDG) and a segmentation network (3D nnUNet), where the former aims to translate source CT images to target multi-sequence MR images which will be combined with the organ masks of their corresponding CT images to train the latter.

The main highlight of OMDG is the comprehensive consideration of training cost and segmentation performance for unsupervised multi-modality domain-adaptive segmentation task. For the former, OMDG benefiting from its characteristic of one-to-multi image translation avoids to train a generator for each pair of CT and certain MR sequence. Compared with one-to-one image translation framework such as CycleGAN, OMDG shows obvious strength in the model parameters, training and inference efficiency. Additionally, although only CT-to-MRI generation results are detailedly analyzed in the previous section, OMDG is essentially a multi-to-multi image translation framework,

Table 8
Quantitative segmentation results of different methods for task 3 in the CHAOS challenge.

		Dice	RAVD	ASSD	MSSD
	SIFA	61.25 ±15.13[a]	103.10 ±81.51[a]	27.34 ±16.58[a]	137.27 ±54.36[a]
I P	CycleGAN	**91.46** **±3.39**	7.77 ±6.17	**2.73** **±1.53**	29.96 ±15.49
	MUNIT	88.60 ±1.96	**6.84** **±5.15**	3.07 ±1.15[a]	**26.06** **±12.18**
	StarGAN v2	78.69 ±17.48[a]	22.58 ±52.17	8.17 ±11.05[a]	50.98 ±37.08[a]
	OMUDA	88.85 ±8.72	10.23 ±13.18	3.06 ±2.73	30.49 ±15.14
	SIFA	85.62 ±5.34[a]	11.57 ±11.22	5.49 ±2.80[a]	53.59 ±18.63[a]
OP	CycleGAN	90.74 ±4.12[a]	13.23 ±7.06[a]	2.79 ±1.71[a]	34.40 ±15.03
	MUNIT	87.19 ±2.63[a]	7.16 ±4.86	3.61 ±1.24[a]	30.81 ±11.02[a]
	StarGAN v2	78.49 ±26.72[a]	20.80 ±21.78[a]	11.92 ±24.17[a]	46.86 ±49.58[a]
	OMUDA	**93.09** **±3.31**	**5.60** **±6.27**	**1.88** **±1.56**	**23.26** **±9.53**
	SOTA (fully)	95.35 ±1.15	3.57 ±2.27	1.26 ±0.83	20.57 ±12.89
	SIFA	55.83 ±15.71[a]	166.93 ±127.99[a]	36.79 ±17.46[a]	168.64 ±59.41[a]
T2	CycleGAN	91.88 ±2.29	**3.94** **±3.18**	2.62 ±1.24	32.61 ±16.86[a]
	MUNIT	89.48 ±2.64[a]	5.24 ±4.67	3.09 ±1.00[a]	29.39 ±10.76
	StarGAN v2	84.88 ±3.61[a]	21.16 ±6.35[a]	4.07 ±1.24[a]	31.68 ±14.28
	OMUDA	**92.19** **±2.17**	4.61 ±3.24	**2.42** **±1.14**	**28.45** **±15.47**
	SOTA (fully)	95.49 ±1.81	2.13 ±1.59	1.39 ±1.01	21.14 ±12.96

mean ± *std*.
[a]Represents that the results achieved by this method are significantly different ($p \leq 0.05$) from those achieved by OMUDA.

Table 9
The correlation of intensity distribution between the real MR images and the synthetic MR images generated by different methods.

	T1 W	EAP	PVP	DP	IP	OP	T2 W	DWI	Avg.
SIFA	0.85	0.91	0.87	0.92	0.80	0.90	0.68	0.72	0.83
CycleGAN	0.81	**0.92**	0.88	0.92	0.63	0.47	0.50	0.35	0.69
MUNIT	0.46	0.48	0.66	0.53	0.80	0.77	**0.78**	**0.95**	0.68
StarGAN v2	0.88	0.91	0.91	0.92	0.50	0.68	0.76	0.53	0.76
OMUDA	**0.93**	**0.92**	**0.92**	**0.95**	**0.96**	**0.97**	0.76	0.93	**0.92**

Table 10
Model parameters and FLOPs of different methods.

	Sub-Nets	Parameters	FLOPs	
			Training	Inference
CycleGAN	G_A	$11.37M \times 9$	$92.72G \times 9$	$46.36G \times 9$
	G_B	$11.37M \times 9$	$92.72G \times 9$	–
	D_A	$2.76M \times 9$	$9.36G \times 9$	–
	D_B	$2.76M \times 9$	$9.36G \times 9$	–
	Total	$28.26M \times 9$	$204.16G \times 9$	$46.36G \times 9$
OMDG	G	$10.76M$	$159.24G$	$286.84G$
	D	$20.85M$	$45.5G$	–
	M	$8.18M$	$16.36M$	8.18M
	E_S	$21.15M$	$33.00G$	–
	Total	$61.60M$	$237.76G$	$286.85G$

G: Generator, D: Discriminator, M: Mapping Net, E_S: Style Encoder.

which means that it is capable of achieving mutual image translation among arbitrary two MRI sequences. In regard to the segmentation performance, although the segmentation performance of OMDG is inferior to that of CycleGAN in a majority of organs (i.e., spleen, kidney and liver) in T1 W and dynamic-contrast enhanced sequences, it is superior to CycleGAN in some small organs, such as gallbladdder and IVC. Moreover, OMDG possesses an apparent advantage over modalities which have more different appearance from the source CT images, e.g., IP, T2 W. Note that due to its advantages in annotation-efficiency, training cost and segmentation performance, this OMUDA framework has been applied in practice and has contributed to the development of SenseCare's first-generation MR liver analysis product (Duan et al., 2020) which has been online in several hospitals. According to the questionnaire results, the segmentation results of about 95% cases meet the clinical requirements and has no need for secondary manual modification by doctors, which paves the way for subsequent lesion detection on individual sequences, joint assessment of lesions by multiple sequences and size measurements of liver, kidney, and spleen resection.

Annotation-free or annotation-efficiency has drawn a lot of attention in the field of medical image processing. In OMUDA, the concept of annotation-efficiency is fully embodied. Besides no annotated organ masks of different MRI sequences is needed in this framework, the annotated organ masks of CT images are almost from public databases, which largely reduce the cost of manual annotation. In addition, OMUDA employs marginal Dice and CE losses to replace ordinary segmentation losses, e.g., Dice loss or CE loss, to address the partial label issue and unify different databases into the same training phase. Combined with the highlights mentioned above, OMUDA shows its strength in unsupervised multi-modality domain-adaptive segmentation task, and can significantly reduce the labor and time cost in some practical scenarios such as the initial phase of product development.

However, there are still some limitations of the OMUDA framework. First, although OMUDA provides a possible method for one-to-multiple unsupervised domain adaptation in the segmentation task, it may fail in some challenging practical applications, e.g., the segmentation of soft tissues or tiny objects. One reason for these failures lies in the different imaging principles of CT and MRI. CT images have much less soft-tissue contrast compared with MRI images, which limits the generation quality of soft tissues in the synthetic MR images. This is an inherent issue when translating CT images to MR images, which cannot be easily solved. In case of segmenting tiny organs, compared with segmenting some relatively large organs, such as liver, spleen and kidney, it has a higher requirement for OMDG in preserving the structural consistency and generation fidelity during image translation. Therefore, further improving the quality of the synthetic images (both in fidelity and anatomical consistency) will become our main focus in the future. Additionally, jointly training the generator by different modalities is one of the highlight of multi-to-multi image translation framework, but this point is not detailedly inquired in OMUDA. Consequently, exploring the advantages caused by joint training of OMDG in generation quality remains to be implemented. Lastly, more attempts will be made to further improve the efficiency of inference.

6. Conclusion

A unified framework called OMUDA for one-to-multiple unsupervised domain adaptation in segmentation of abdominal organs was proposed in our study, where a model trained from a single source dataset is adapted to multiple target domains simultaneously and efficiently without extra annotations. In OMDG, generator refactoring and style constraint are adopted for better anatomical consistency and domain-aliasing reduction respectively. OMUDA achieves an average DSC of 85.51% on the in-house test set and an average DSC of 82.66% on the AMOS22 dataset, slightly lower than CycleGAN (85.66% and 83.40%), However, it achieves an average DSC of 90.61% across IP, OP, T2 W and DWI in the internal test set and 91.38% on the CHAOS dataset (CycleGAN: 82.05% and 91.36%), showing its better segmentation performance in modalities which have even more different appearance

from the source modality. Furthermore, compared with CycleGAN, our OMDG reduces FLOPs by about 87% and 30% in the training and inference stage respectively, demonstrating its gains in the training and inference efficiency. In our future work, we will focus on improving the quality of some challenging objects in the synthetic images, such as tiny organs and soft tissues, which are not detailly investigated in our current work.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential

Acknowledgments

This work was supported by the National Natural Science Foundation of China [62271115], the Key Research and Development Program of Zhejiang Province [2019C03064], the Action Plan of Shanghai Science and Technology Commission [21SQBS02300] and the Shanghai Xuhui District Hospital and Enterprise Cooperation Plan [2021-015].

Appendix. Hyper-parameters in comparison study

SIFA: We reimplemented SIFA in PyTorch based on the TensorFlow version.[15] All settings and hyper-parameters of the networks and losses used in our experiment were same as those in the open source code.[16]

CycleGAN In our experiment, ResNet_9block with 2 down-sampling operations and PatchGAN discriminator with 4 convolution layers[17] were chosen as the architectures of the generator and the discriminator respectively. L1 loss was employed as a cyclic loss to restrict the consistency between the original image and the reconstructed image, and mean square Error (MSE) loss was used to discriminate the authenticity of generated images. Furthermore, we utilized the identity loss to facilitate the training of the generators. The weights for these three losses were set to 10, 1, 5 respectively. The batch size was set to 8 during training. In the segmentation stage, all settings and hyper-parameters are same as those in OMUDA.

MUNIT The architectures of the networks are same as which in.[18] The weights for the adversarial loss, the image reconstruction loss, style reconstruction loss, content reconstruction loss, cyclic consistency loss and domain-invariant perceptual loss were 1, 10, 1, 10 and 0 respectively. As in the above experiments, the batch size was set to 8. As for the training of nnUNet, all settings and hyper-parameters are same as those in OMUDA, which are self-configured by the nnUNet framework.

StarGAN v2 No changes were made to the network architectures in.[19] The latent dimension, hidden dimension and style dimension in the style encoder were set to 16, 256 and 24 in our experiment. The weights for R1 regression, cyclic consistency loss and style reconstruction loss, diversity sensitive loss were 1, 1, 1 and 2. No high-pass filtering was used in this comparative experiment. All settings and hyper-parameters in the segmentation stage were self-planned by the nnUNet framework.

[15] https://github.com/cchen-cc/SIFA

[16] https://github.com/JianghaoWu/SIFA-pytorch.git

[17] https://github.com/harveerar/PMB_2020_Self_Derived_OAD_Segmentor/tree/master/PMB_attention_code

[18] https://github.com/NVlabs/MUNIT

[19] https://github.com/clovaai/stargan-v2

References

Avants, B.B., Tustison, N., Song, G., et al., 2009. Advanced normalization tools (ANTS). Insight J. 2 (365), 1–35.

Bobo, M.F., Bao, S., Huo, Y., Yao, Y., Virostko, J., Plassard, A.J., Lyu, I., Assad, A., Abramson, R.G., Hilmes, M.A., et al., 2018. Fully convolutional neural networks improve abdominal organ segmentation. In: Medical Imaging 2018: Image Processing, vol. 10574, International Society for Optics and Photonics, p. 105742V.

Chartrand, G., Cresson, T., Chav, R., Gotra, A., Tang, A., De Guise, J.A., 2016. Liver segmentation on CT and MR using Laplacian mesh optimization. IEEE Trans. Biomed. Eng. 64 (9), 2110–2121.

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.

Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., Choo, J., 2018. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8789–8797.

Choi, Y., Uh, Y., Yoo, J., Ha, J.W., 2020. Stargan v2: Diverse image synthesis for multiple domains. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR.

Conze, P.H., Kavur, A.E., Cornec-Le Gall, E., Gezer, N.S., Le Meur, Y., Selver, M.A., Rousseau, F., 2021. Abdominal multi-organ segmentation with cascaded convolutional and adversarial deep networks. Artif. Intell. Med. 117, 102109.

Dou, Q., Chen, H., Jin, Y., Yu, L., Qin, J., Heng, P.A., 2016. 3D deeply supervised network for automatic liver segmentation from CT volumes. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 149–157.

Du, L., Tan, J., Yang, H., Feng, J., Xue, X., Zheng, Q., Ye, X., Zhang, X., 2019. Ssf-dan: Separated semantic feature based domain adaptation network for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 982–991.

Duan, Q., Wang, G., Wang, R., Fu, C., Li, X., Gong, M., Liu, X., Xia, Q., Huang, X., Hu, Z., et al., 2020. SenseCare: A research platform for medical image informatics and interactive 3D visualization. arXiv preprint arXiv:2004.07031.

Ge, Y., Chen, D., Li, H., 2020. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. arXiv preprint arXiv:2001.01526.

Gholami, B., Sahu, P., Rudovic, O., Bousmalis, K., Pavlovic, V., 2020. Unsupervised multi-target domain adaptation: An information theoretic approach. IEEE Trans. Image Process. 29, 3993–4002.

Gibson, E., Giganti, F., Hu, Y., Bonmati, E., Bandula, S., Gurusamy, K., Davidson, B., Pereira, S.P., Clarkson, M.J., Barratt, D.C., 2018. Automatic multi-organ segmentation on abdominal CT with dense v-networks. IEEE Trans. Med. Imaging 37 (8), 1822–1834.

Heinrich, M.P., Jenkinson, M., Bhushan, M., Matin, T., Gleeson, F.V., Brady, M., Schnabel, J.A., 2012. MIND: Modality independent neighbourhood descriptor for multi-modal deformable registration. Med. Image Anal. 16 (7), 1423–1435.

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., Belongie, S., 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In: 2017 IEEE International Conference on Computer Vision. ICCV.

Huang, X., Liu, M.Y., Belongie, S., Kautz, J., 2018. Multimodal unsupervised image-to-image translation. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 172–189.

Isensee, F., Jäger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H., 2019a. Automated design of deep learning methods for biomedical image segmentation. arXiv preprint arXiv:1904.08128.

Isensee, F., Petersen, J., Kohl, S.A., Jäger, P.F., Maier-Hein, K.H., 2019b. Nnu-net: Breaking the spell on successful medical image segmentation. 1, (1–8), p. 2, arXiv preprint arXiv:1904.08128.

Isobe, T., Jia, X., Chen, S., He, J., Shi, Y., Liu, J., Lu, H., Wang, S., 2021. Multi-target domain adaptation with collaborative consistency learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8187–8196.

Jiang, X., Lao, Q., Matwin, S., Havaei, M., 2020. Implicit class-conditioned domain alignment for unsupervised domain adaptation. In: International Conference on Machine Learning. PMLR, pp. 4816–4827.

Karasawa, K., Oda, M., Kitasaka, T., Misawa, K., Fujiwara, M., Chu, C., Zheng, G., Rueckert, D., Mori, K., 2017. Multi-atlas pancreas segmentation: Atlas selection based on vessel structure. Med. Image Anal. 39, 18–28.

Kart, T., Fischer, M., Küstner, T., Hepp, T., Bamberg, F., Winzeck, S., Glocker, B., Rueckert, D., Gatidis, S., 2021. Deep learning-based automated abdominal organ segmentation in the UK biobank and German national cohort magnetic resonance imaging studies. Invest. Radiol. 56 (6), 401–408.

Kavur, A.E., Gezer, N.S., Barış, M., Aslan, S., Conze, P.H., Groza, V., Pham, D.D., Chatterjee, S., Ernst, P., Özkan, S., et al., 2021. CHAOS challenge-combined (CT-MR) healthy abdominal organ segmentation. Med. Image Anal. 69, 101950.

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.

Liu, M.Y., Breuel, T., Kautz, J., 2017. Unsupervised image-to-image translation networks. Adv. Neural Inf. Process. Syst. 30.

Liu, M.Y., Tuzel, O., 2016. Coupled generative adversarial networks. Adv. Neural Inf. Process. Syst. 29.

Luo, X., Liao, W., Xiao, J., Chen, J., Song, T., Zhang, X., Li, K., Metaxas, D.N., Wang, G., Zhang, S., 2022. WORD: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from CT image. Med. Image Anal. 82, 102642.

Ma, J., Zhang, Y., Gu, S., Zhu, C., Ge, C., Zhang, Y., An, X., Wang, C., Wang, Q., Liu, X., et al., 2021. Abdomenct-1k: Is abdominal organ segmentation a solved problem. IEEE Trans. Pattern Anal. Mach. Intell..

Pham, D.D., Dovletov, G., Warwas, S., Landgraeber, S., Jäger, M., Pauli, J., 2019. Deep learning with anatomical priors: Imitating enhanced autoencoders in latent space for improved pelvic bone segmentation in MRI. In: 2019 IEEE 16th International Symposium on Biomedical Imaging. ISBI 2019, IEEE, pp. 1166–1169.

Roth, H.R., Oda, H., Hayashi, Y., Oda, M., Shimizu, N., Fujiwara, M., Misawa, K., Mori, K., 2017. Hierarchical 3D fully convolutional networks for multi-organ segmentation. arXiv preprint arXiv:1704.06382.

Roth, H., Oda, M., Shimizu, N., Oda, H., Hayashi, Y., Kitasaka, T., Fujiwara, M., Misawa, K., Mori, K., 2018. Towards dense volumetric pancreas segmentation in CT using 3D fully convolutional networks. In: Medical Imaging 2018: Image Processing, vol. 10574, International Society for Optics and Photonics, p. 105740B.

Savenije, M.H., Maspero, M., Sikkes, G.G., van der Voort van Zyp, J., Kotte, T., Alexis, N., Bol, G.H., van den Berg, T., Cornelis, A., et al., 2020. Clinical implementation of MRI-based organs-at-risk auto-segmentation with convolutional networks for prostate radiotherapy. Radiat. Oncol. 15 (1), 1–12.

Sharma, A., Hamarneh, G., 2019. Missing MRI pulse sequence synthesis using multi-modal generative adversarial network. IEEE Trans. Med. Imaging 39 (4), 1170–1183.

Shi, G., Xiao, L., Chen, Y., Zhou, S., 2020. Marginal Loss and Exclusion Loss for Partially Supervised Multi-Organ Segmentation.

Sohn, K., 2016. Improved deep metric learning with multi-class N-pair loss objective. In: Neural Information Processing Systems.

Tang, M., Djelouah, A., Perazzi, F., Boykov, Y., Schroers, C., 2018. Normalized cut loss for weakly-supervised cnn segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1818–1827.

Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., Darrell, T., 2014. Deep domain confusion: Maximizing for domain invariance. arXiv preprint arXiv:1412.3474.

Wu, J., Gu, R., Dong, G., Wang, G., Zhang, S., 2022. FPL-uda: Filtered pseudo label-based unsupervised cross-modality adaptation for vestibular schwannoma segmentation. In: 2022 IEEE 19th International Symposium on Biomedical Imaging. ISBI, IEEE, pp. 1–5.

Wu, W., Zhou, Z., Wu, S., Zhang, Y., 2016. Automatic liver segmentation on volumetric CT images using supervoxel-based graph cuts. Comput. Math. Methods Med. 2016.

Zhou, X., Ito, T., Takayama, R., Wang, S., Hara, T., Fujita, H., 2016. Three-dimensional CT image segmentation by combining 2D fully convolutional network with 3D majority voting. In: Deep Learning and Data Labeling for Medical Applications. Springer, pp. 111–120.

Zhu, J.Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2223–2232.