

A novel 3D unsupervised domain adaptation framework for cross-modality medical image segmentation

Kai Yao, Zixian Su, Kaizhu Huang, Xi Yang, Jie Sun, Amir Hussain, Frans Coenen

Abstract— We consider the problem of volumetric (3D) unsupervised domain adaptation (UDA) in cross-modality medical image segmentation, aiming to perform segmentation on the unannotated target domain (e.g. MRI) with the help of labeled source domain (e.g. CT). Previous UDA methods in medical image analysis usually suffer from two challenges: 1) they focus on processing and analyzing data at 2D level only, thus missing semantic information from the depth level; 2) one-to-one mapping is adopted during the style-transfer process, leading to insufficient alignment in the target domain. Different from the existing methods, in our work, we conduct a first of its kind investigation on multi-style image translation for complete image alignment to alleviate the domain shift problem, and also introduce 3D segmentation in domain adaptation tasks to maintain semantic consistency at the depth level. In particular, we develop an unsupervised domain adaptation framework incorporating a novel quartet self-attention module to efficiently enhance relationships between widely separated features in spatial regions on a higher dimension, leading to a substantial improvement in segmentation accuracy in the unlabeled target domain. In two challenging cross-modality tasks, specifically brain structures and multi-organ abdominal segmentation, our model is shown to outperform current state-of-the-art methods by a significant margin, demonstrating its potential as a benchmark resource for the biomedical and health informatics research community.¹

Index Terms— Cross-modality learning, Image segmentation, Style transfer, Unsupervised domain adaptation.

I. INTRODUCTION

RECENT years have witnessed the bloom of deep convolutional neural networks (CNNs) in medical image processing [1]–[3]. However, well-trained deep models usually perform poorly in real scenarios due to the severe data distribution difference between training and test sets caused

Kai Yao and Zixian Su contribute equally to this work.

Kai Yao and Zixian Su are with both University of Liverpool and School of Advanced Technology, Xi’an Jiaotong-Liverpool University.

Kaizhu Huang is with Institute of Applied Physical Sciences and Engineering, Duke Kunshan University, Kunshan, Jiangsu, 215316, China.

Xi Yang and Jie Sun are with the School of Advanced Technology, Xi’an Jiaotong-Liverpool University, Suzhou, Jiangsu, 215000, China.

Amir Hussain is with the School of Computing, Edinburgh Napier University, Edinburgh, EH11 4BN, United Kingdom.

Frans Coenen is with University of Liverpool, United Kingdom.

Correspondence: kaizhu.huang@dukekunshan.edu.cn and xi.yang01@xjtlu.edu.cn.

¹Code is available at: <https://github.com/Kaiseem/DAR-UNet>

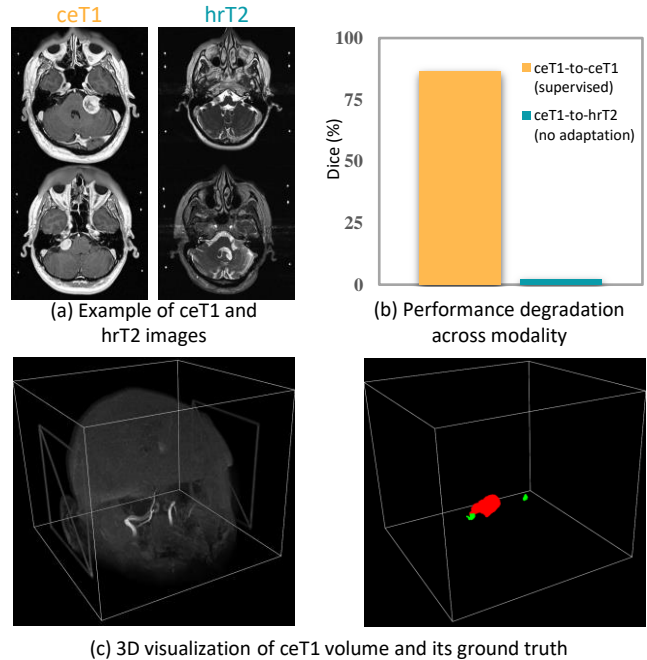


Fig. 1. Illustration of challenges from cross-modality medical images. (a) 2D visual comparison between ceT1 and hrT2 images. (b) Comparison of the models trained on ceT1 images while evaluated on ceT1 images and hrT2 images using Dice score, denoted as “supervised” and “no adaptation”, respectively. (c) 3D visualization of ceT1 image and its corresponding label.

by different imaging modalities, scanning protocols, and/or demographic properties. For instance, the contrast-enhanced T1 (ceT1) Magnetic Resonance Imaging (MRI) scans and high-resolution T2 (hrT2) scans are commonly used for the follow-up and treatment planning of vestibular schwannoma (VS), in which two key brain structures, the tumour and cochlea, are expected to be segmented in clinical practice [4]. However, a large visual appearance variation can be observed between ceT1 and hrT2 scans, as shown in Figure 1 (a). A model purely trained on ceT1 images cannot directly generalize well on hrT2 images due to the distribution shift between these two modalities (seen in Figure 1 (b)).

To reduce the performance degradation across different modalities, a straightforward way is to fine-tune the model pre-trained on source data using labeled target data [5], [6].

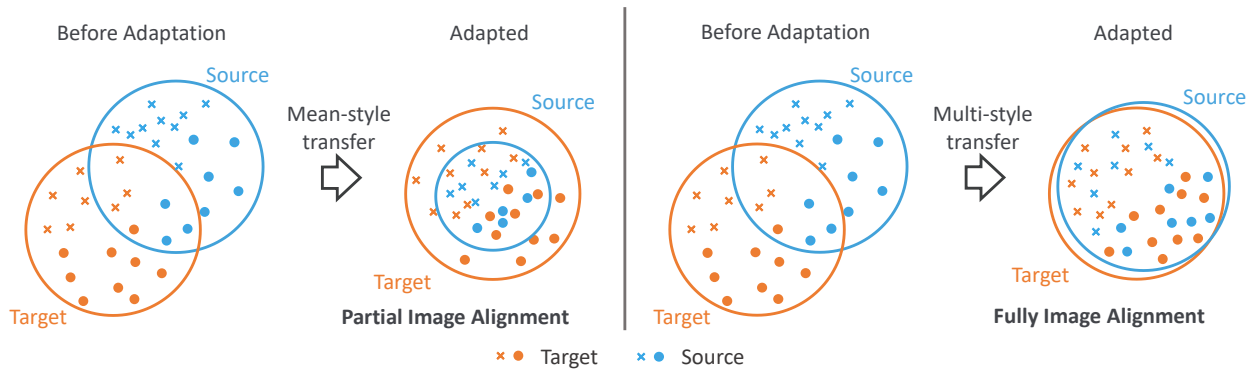


Fig. 2. (Left) The transferred source images with mean-style of target domain might not reflect the true distribution of target domain, (Right) while using multiple samples' styles to transfer the source images can generate a target-like distribution of transferred images.

However, this method is often not suitable in real situations especially in medical images. When we exploit pretrained models in a practical scenario, annotated target datasets are often limited due to the tedious labeling process and privacy concerns. Instead, unsupervised domain adaptation (UDA) is more attractive and feasible, where the ground truth in the target domain is not required.

Most current UDA-based methods in medical image analysis can be summarized from two perspectives. One is image alignment, which transforms the image style from source domain to target domain using unpaired image-to-image translation. Another is feature alignment, aiming to extract domain-invariant features via adversarial training. Although the existing UDA-based methods in medical image analysis have made large progress, drawbacks still remain in them.

First, they tend to ignore the 3D information inherently available in medical images, but choose to cut the medical volumes into slices and use 2D networks instead, where abundant semantic information from depth level is left aside in their settings [7]. These methods may work on tasks with small domain adaptation and segmentation difficulty, where the random slice training set can provide enough information for segmentation and domain adaptation. However, for more challenging tasks with larger segmentation difficulty, the absence of semantic consistency in depth channel would degrade the overall performance in the target domain. In the case of vestibular schwannoma segmentation as mentioned before, the two structures, the tumour and cochlea, only take up 0.15% volume of the whole sample (seen in Figure 1 (c)), which poses a great challenge to the segmentation network. If only 2D slices are adopted, most training samples only contain background information. Such imbalance issue between the foreground and background would result in the failure of segmentation network. Moreover, from the viewpoint of data analysis, using neural networks to process a full 3D image directly rather than 2D slices is more reasonable and interpretable as the former is closer to what the human eyes are exposed to.

Second, an overly simplified assumption for style transfer, which is a commonly used strategy of image alignment [8], [9], is adopted in most previous work, where they model this process as a deterministic one-to-one mapping with a mean

style of each domain. Namely, for a given input source image, such methods can only synthesize one rigid output with the mean style of target domain. However, there exists an inherent intra-domain variance in medical image datasets, meaning that the image styles in a single domain are quite different. The mean-style transformation would lead to the so-called partial image alignment as shown in Figure 2 (Left), which will undoubtedly weaken the model generalizability on edge samples (samples far away from center point in style space) in the target domain. To alleviate this shortcoming, some work chooses to combine feature alignment in the following stage for further alignment, which somehow lacks controllability as an intermediate output in the high dimension space.

In this paper, to better solve the above problems, we propose a novel 3D unsupervised domain adaptation framework. Specifically, a generative adversarial network (GAN) is first trained for content-style disentangled cross-domain image-to-image translation. Unlike previous efforts that encode domains into a common feature space, our GAN extracts the domain-invariant features as content and domain-specific features as style separately, enabling a target-like image generation with diverse styles. This diverse generation method allows complete image alignment in target domain as shown in Figure 2 (Right) to better reduce domain shift from image level; it can also avoid further feature alignment in 3D segmentor, which may introduce additional computational overhead. After that, the synthesized target-like images (volumes) are utilized for training our proposed 3D dual attention residual U-Net (DAR-UNet) segmentor. We implement a 3D *voxel-wise Attention Module* (VAM) in the decoder part of the segmentation subnet to focus on the essential areas of the feature maps that can prioritize the effective areas for segmentation. Meanwhile, a *Quartet Attention Module* (QAM) is adopted in each residual block to stress the adjacent semantic information between slices, which can effectively capture the semantic consistency from the depth level.

In summary, the key contributions of this paper are as follows:

- We propose DAR-UNet architecture, a novel 3D semantic segmentation neural network that takes advantages of two attention modules, VAM and QAM, to capture spatial semantic information from various feature levels.

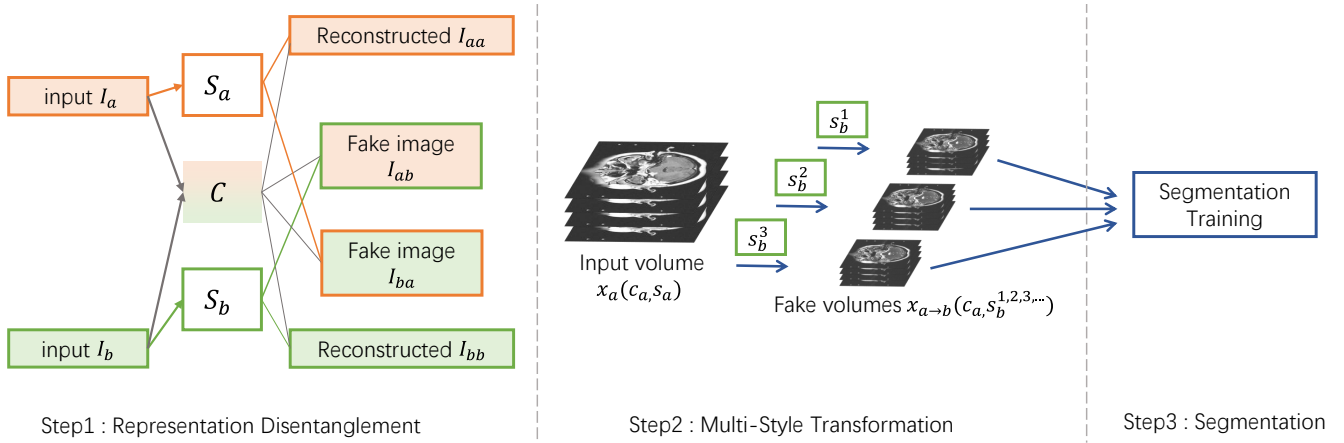


Fig. 3. General process of our proposed framework. We first factorize the source domain images and target domain images into content and style via disentangled generative adversarial network. Then, for each input source volume, we generate target-like volumes with diverse styles. Finally, we train our DAR-UNet with the transferred volumes.

- We investigate the feasibility of multi-style transfer strategy for image alignment in cross-modality medical image segmentation tasks, which shows impressive performance in reducing distribution shift.
- DAR-UNet outperforms the state-of-the-art methods by a large margin on two benchmarks, i.e., vestibular schwannoma dataset and abdominal multi-organ datasets.

II. RELATED WORK

A. Unsupervised Domain Adaptation

In domain adaptation, training images are denoted as the source domain, while test images are termed as the target domain. They tend to have similar content but different style distributions. This phenomenon is known as domain shift in machine learning. In medical image analysis, due to the prevalence of varying hospital modalities and diverse hospital populations, domain shift is even more serious compared with conventional common data [10], [11]. Previous studies [12] have revealed that the error of the model on the test set is usually proportional to the domain shift from the training set. In the case of large domain shift as commonly seen in clinical practice, models trained on the source set tend to perform poorly on target domain. Therefore, how to transfer the knowledge from source domain to target domain in the field of medical images is a critical problem. While relabeling in target domain is labor-consuming and time-costly, unsupervised domain adaptation (UDA) requiring no annotation has attracted much interest as a promising and feasible alternative.

Recent research adopts adversarial learning to tackle the domain shift problem from different aspects, including pixel-level alignment [13], feature-level alignment [14], and the joint learning methods [15]. Pixel-level alignment, also called image transformation, adopts an image-to-image generative model to convert the style of source domain into that of target domain while keeping its content unchanged. Cooperated with a discriminator, the style-transferred images are designed to be indistinguishable with the target domain data. Feature alignment, termed latent feature space transformation, aligns the distributions across domains in the feature space rather

than the image space, in order to mitigate the domain shift. In this process, the generator does not need to generate new samples anymore; instead, it plays the role of feature extraction to learn constantly the common characteristics of the data.

B. Representation Disentangling

Representation disentangling in different domains aims to model factors of data variation implicitly, in which the representations with small variation can be roughly regarded as domain-invariant features, while those with large disparity are domain-specific. Some previous work [16] takes unpaired data to factorize representations into the content and attribute space to produce diverse outputs. However, they only focus on learning disentangled representation of images in a single domain and cannot be easily extended to describe cross-domain data. Recently, Liu *et al.* have proposed E-CDRD [17], the first framework that addresses cross-domain disentangled representation with only supervision from single-domain data. Later, UFDN [18], a unified deep learning framework, is presented to learn domain-invariant representations from data across multiple domains in an unsupervised setting. After that, many excellent approaches have investigated disentangled representations of content and style for image-to-image translation. Although it is hard to define the content/style explicitly in different domains, it is assumed that two domains share the same content but different styles. Moreover, domain-invariant features are taken as the content revealing the spatial structure of the segmentation target, while the rendering of the structure is regarded as styles. Specifically, in DRIT [19], features of each domain are disentangled into either domain-invariant features (DIFs) or domains-specific features (DSFs), which we also follow in this paper. With the disentangled representations, image-to-image style translation is performed by swapping the content and the style extractions.

C. Attention Module

The attention mechanism emerges as improvement over the encoder-decoder-based translation system in natural language

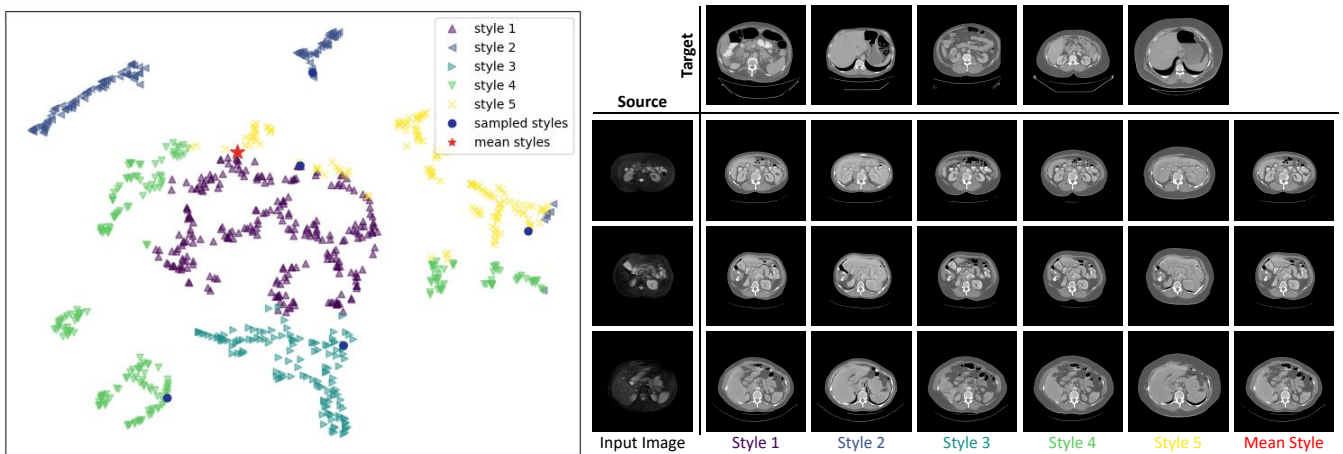


Fig. 4. Visualization of multi-style image transformation via our proposed content-style disentangled GAN on abdominal multi-organ datasets. (Left) Visualization of the target domain style space using T-SNE [20]. The styles are clustered into 5 classes using K-Means [21], and the blue triangles represent the sampled styles which are used to generate the target-like images. Red star indicates the mean of all styles. (Right) The generated target-like images with diverse styles. Each image is generated with the content of source image in the left column and the style of sampled target image in the above row.

processing (NLP) to solve the long-range dependency problem of RNN/LSTMs [22], [23]. It is then extended to computer vision [24] so that global dependencies between input and output can be well learned with more attentions on more important locations. One pioneering work is residual attention network (RAN) [25]. As the first attempt in visual recognition to use a non-local operation, RAN can capture long-range dependencies though its computational cost could be very high. Subsequently, Zhang et al. creatively propose SAGAN combining self-attention with GAN [26]. This method allows attention-driven, long-range dependency modeling for image generation tasks efficiently. While most existing works investigate the relationships in the 2D plane space of input, triplet attention is proposed in [27] that comprises of three branches, each responsible for capturing cross-dimension information between the spatial dimensions and channel dimensions of the input. Inspired by their work, we propose a quadruple attention module that incorporates the information from depth channel thanks to the available 3D information in the datasets.

III. MAIN METHOD

Our main idea is to rephrase the 3D UDA problem as an image transformation problem and a segmentation problem, as shown in Figure 3. Concretely, our system consists of a 2D translation network and a 3D segmentation network. The translation network disentangles the content and style of the 2D slices from both domains. The translated images are produced by randomly sampling one domain's style code and recombining it with the content code. Then, the translated volumes are utilized for training the segmentation network under supervision. In the following, we will first describe the translation networks to minimize the domain gap. Afterwards, we will detail the employed segmentor for 3D segmentation.

A. Translation Network for Content-Style Disentangling

Let $x_a \in \mathbb{R}^{D \times W \times H}$ be a given source domain greyscale volume of domain \mathcal{X}_a , $y_a \in \mathbb{R}^{C \times D \times W \times H}$ be the correspond-

ing 3D labels \mathcal{Y}_a , and $x_b \in \mathbb{R}^{D \times W \times H}$ be a given target domain greyscale volume of domain \mathcal{X}_b , where C, D, W, H indicate the class number, depth, width, and height, respectively. $I_a \in \mathbb{R}^{W \times H}$ and $I_b \in \mathbb{R}^{W \times H}$ are the slices of the volumes \mathcal{X}_a and \mathcal{X}_b along the Z -axis. Our aim is to achieve promising performance in target domain by applying the model trained on the labelled source domain. To this end, we develop a novel 2D content-style disentangled translation network in this paper. Specifically, we assume that the latent space of images $I \in \mathcal{I}$ can be factorized into a content code $c \in \mathcal{C}$ and a style code $s \in \mathcal{S}$. Our network consists of one shared content encoder $\mathcal{G}_{enc,c}$, two domain-specific style encoder $\mathcal{G}_{enc,s_a}, \mathcal{G}_{enc,s_b}$, one shared decoder \mathcal{G}_{dec} , one content discriminator \mathcal{D}_c , and two image discriminators $\mathcal{D}_a, \mathcal{D}_b$. The encoders factorize the input images into content representations and style representations, while the decoder reconstructs the image from content representation by injecting the style representations via Adaptive Instance Normalization (AdaIN) layers [28]. We introduce reconstruction losses in content-level, style-level, and pixel-level to maintain the semantic information. In particular, pixel-level adversarial training is used to learn cross-domain image-to-image translation, while content-level adversarial training is utilized to ensure a fine content feature alignment along with the shared content encoder. The detailed descriptions about reconstruction losses and adversarial losses of the translation network are provided in the following.

1) **Reconstruction Loss:** To ensure a bijective mapping between \mathcal{I} and $\{\mathcal{C}, \mathcal{S}\}$ and preserve content and style information of both domains, three reconstruction losses are employed at pixel-level, content-level, and style-level, respectively. We first perform the same-domain translation to reconstruct the input source image I_a by passing it through the shared content encoder $\mathcal{G}_{enc,c}$, specific style encoder \mathcal{G}_{enc,s_a} , and the shared decoder \mathcal{G}_{dec} . The image reconstruction loss is optimized,

$$\mathcal{L}_{rec}^{I_a} = \mathbb{E}_{I_a \sim \mathcal{I}_a} \|I_a - \mathcal{G}_{dec}(\mathcal{G}_{enc,c}(I_a), \mathcal{G}_{enc,s_a}(I_a))\|_1.$$

The content-level loss is engaged during the cross-domain

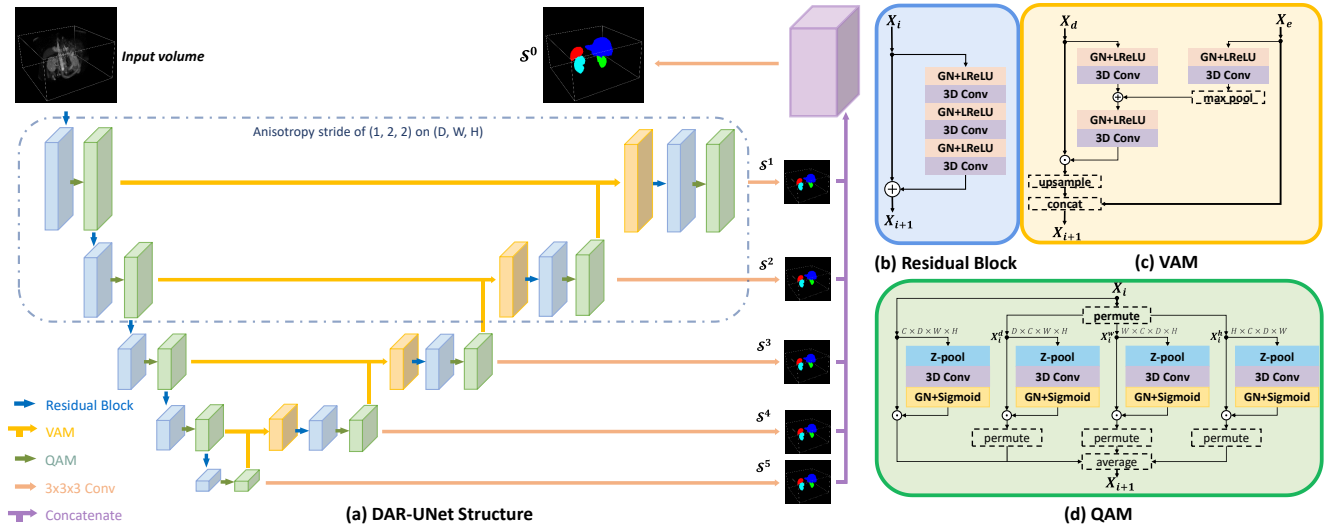


Fig. 5. Overview of our proposed DAR-UNet for 3D segmentation. (a) The hybrid architecture of the proposed DAR-UNet. The proposed VAM enhances the encoder-guided feature map, while the efficient QAM captures the dependencies across dimensions. (b) Pre-activation residual block. (c) The structure of voxel-wise attention module (VAM). (d) The structure of quartet attention module (QAM).

translation. Given a content code c_a encoded from source domain, and a style code s_b randomly encoded from target domain, we should be able to reconstruct it after decoding and encoding,

$$\mathcal{L}_{rec}^{c_a} = \mathbb{E}_{c_a \sim \mathcal{C}_a, s_b \sim \mathcal{S}_b} \|c_a - \mathcal{G}_{enc.c}(\mathcal{G}_{dec}(c_a, s_b))\|_1.$$

Similarly, we can also formulate the style-level loss of translation networks as below,

$$\mathcal{L}_{rec}^{s_a} = \mathbb{E}_{s_a \sim \mathcal{S}_a, c_b \sim \mathcal{C}_b} \|s_a - \mathcal{G}_{enc.s_b}(\mathcal{G}_{dec}(c_b, s_a))\|_1.$$

The target domain images go to the same path as above-mentioned. Combining them together, we obtain the reconstruction loss for the translation network,

$$\begin{aligned} \mathcal{L}_{rec} = & \lambda_I (\mathcal{L}_{rec}^{I_a} + \mathcal{L}_{rec}^{I_b}) + \lambda_c (\mathcal{L}_{rec}^{c_a} + \mathcal{L}_{rec}^{c_b}) \\ & + \lambda_s (\mathcal{L}_{rec}^{s_a} + \mathcal{L}_{rec}^{s_b}), \end{aligned}$$

where λ_I , λ_c , and λ_s are the hyper-parameter weights to balance different losses.

2) Adversarial Loss: In the cross-domain translation, the generator is trained to synthesize target-like translated images to fool the discriminator, while the discriminator is trained to distinguish from the real images and the translated images. Under such training process, the distribution of translated images can be matched to the target data distribution. The optimization function is shown as follows

$$\begin{aligned} \mathcal{L}_{GAN}^{I_a} = & \mathbb{E}_{c_b \sim \mathcal{C}_b, s_a \sim \mathcal{S}_a} [\log(1 - \mathcal{D}_a(\mathcal{G}_{dec}(c_b, s_a)))] \\ & + \mathbb{E}_{I_a \sim \mathcal{I}_a} [\log(\mathcal{D}_a(I_a))]. \end{aligned}$$

In order to further align the content representations from two domains, we conduct content-level adversarial losses as below

$$\mathcal{L}_{GAN}^c = \mathbb{E}_{c_a \sim \mathcal{C}_a} [\log(1 - \mathcal{D}_c(c_a))] + \mathbb{E}_{c_b \sim \mathcal{C}_b} [\log(\mathcal{D}_c(c_b))].$$

Taking them together, we obtain the adversarial loss for the translation network,

$$\mathcal{L}_{GAN} = \mathcal{L}_{GAN}^{I_a} + \mathcal{L}_{GAN}^{I_b} + \mathcal{L}_{GAN}^c.$$

We jointly train the encoders, decoders, and discriminators to optimize the final objective, which is a weighted sum of the adversarial loss and the reconstruction loss terms

$$\begin{aligned} \min_{\mathcal{G}} \max_{\mathcal{D}} \mathcal{L}(\mathcal{G}_{enc.c}, \mathcal{G}_{enc.s_a}, \mathcal{G}_{enc.s_b}, \mathcal{G}_{dec}, \mathcal{D}_a, \mathcal{D}_b, \mathcal{D}_c) \\ = \mathcal{L}_{GAN} + \mathcal{L}_{rec}. \end{aligned}$$

B. Segmentation Network with Dual Attention

Before the training of segmentation network, we first collect all the style representations of target domain images, and cluster them into n_k classes with K-Means [21]. For a given source domain volume x_a , one style s_b is randomly selected from each clustered class of styles to generate the translated source domain volume $x_{a \rightarrow b}$ slice by slice along the Z-axis, as shown in Figure 4. After the process of multi-style generation, the synthesized volumes are leveraged to train the segmentation net. To maximize the semantic mining in such meaningful data, we propose DAR-UNet which is adapted from the classic 2D ResU-Net [29], [30] to volumetric data by replacing part of the 2D convolutions with 3D convolutions, providing accurate segmentation masks and efficient computation. The detailed descriptions about the components of DAR-UNet are provided in the following subsection.

1) Voxel-wise Attention Module: In the U-Net like architecture, skip connections are adopted to pass the multi-scale encoder feature maps into decoder, followed by a direct concatenation to fuse the features. In order to make the decoder to focus on the essential areas of the feature maps, we take advantage of the attention mechanism which is widely used in computer vision to enhance the quality of features. To be specific, we add a *Voxel-wise Attention Module* (VAM) in each scale of the decoder, as shown in Figure 5 (c). X_e represents the feature from encoder, while X_d denotes the feature from decoder. A voxel-wise attention map is generated with X_e and X_d to enhance the certain area of decoder feature.

2) *Quartet Attention Module*: Most of the existing attention modules are designed to capture the dependencies of channels (C) and/or spatial location (W, H). Recent works [31] demonstrate the potential of capturing the inter-dimensional dependencies to improve networks with a low computation cost. Following their work, we implement a *Quartet Attention Module* (QAM) that captures dependencies between the (D, H, W) , (C, H, W) , (C, D, W) and (C, D, H) dimensions of the input tensor respectively. As shown in Figure 5 (d), this quartet attention module works in a four-branch way to output a refined tensor that better captures the global as well as the local dependencies on higher dimensions. To be specific, given an input tensor $\mathcal{X} \in R^{C \times D \times H \times W}$, we first permute the dimensions of the input tensor to get $\mathcal{X}^d \in R^{D \times C \times H \times W}$, $\mathcal{X}^h \in R^{H \times C \times D \times W}$ and $\mathcal{X}^w \in R^{W \times C \times D \times H}$. Then, Z-pool operation is conducted on the aforementioned four tensors to reduce the zeroth dimension. The Z-pool layer is written as

$$\text{Z-pool}(\mathcal{X}) = [\text{MaxPool}_{0d}(\mathcal{X}), \text{AvgPool}_{0d}(\mathcal{X})].$$

After the reduction of the zero-dimension, four standard three-dimensional convolutional layers are utilized respectively, followed by a sigmoid activation function to calculate the corresponding attention maps. Then, the attention maps are subsequently applied to the tensors before Z-pool. Finally, the refined tensors are rearranged to their original dimension (C, D, H, W) and averaged to output final result. To summarize, the whole process to obtain the refined attention-applied tensor y from the quartet attention can be described as:

$$y = \frac{1}{4} (\overline{\mathcal{X}^d \sigma(\psi_1(\text{Z-pool}(\mathcal{X}^d)))} + \overline{\mathcal{X}^w \sigma(\psi_2(\text{Z-pool}(\mathcal{X}^w)))} + \overline{\mathcal{X}^h \sigma(\psi_3(\text{Z-pool}(\mathcal{X}^h)))} + \overline{\mathcal{X} \sigma(\psi_4(\text{Z-pool}(\mathcal{X})))}),$$

where σ is the sigmoid activation function; ψ_i represents the standard three-dimensional convolutional layers of size k in each branch of quartet attention module; (\cdot) represents the permutation operation to (C, D, H, W) .

3) *Anisotropic Resolution Network*: Many medical images are captured and collected slice-by-slice, resulting in a high in-plane resolution but low through-plane resolution. Traditional isotropic 3D CNNs require the spacing normalized images with isotropic 3D resolution to avoid the imbalance problem of receptive field along each axis [7]. However, since the ratio of the through-plane and in-plane resolution in many medical datasets is about 3-8, isotropic restoration of those images may introduce useless information and storage space redundancy. To this end, we utilize an anisotropic design to reduce the memory cost. Specifically, the main structure of our proposed DAR-UNet follows the typical encoder and decoder design of U-Net, which contains five levels of convolution. As shown in Figure 5, the first two levels and the other three levels use the stride of (1,2,2) and (2,2,2) on (Depth, Width, Height), respectively. With this design, our network can train and infer with the images which have 4 times through-plane resolution than in-plane resolution, enabling a significant reduction of memory cost without performance decline.

4) *Segmentation Loss*: Since the source domain data are annotated, we can train the segmentor \mathcal{S} with the transferred source domain images $x_{a \rightarrow b}$ and their corresponding labels

y_a . We engage a deep supervision strategy [32] to alleviate the potential gradient vanishing problem during training. As shown in Figure 5, one auxiliary prediction branch is applied at each level, while the primary branch merges the total five auxiliary branch features to generate the final prediction. All the prediction branches are optimized with the segmentation loss. We employ a sum of soft Dice loss [33] along with Focal loss [34] to overcome the imbalance issue between the foreground and background to train the segmentor:

$$\mathcal{L}_{seg}(\mathcal{S}) = \mathcal{L}_{Dice}(\mathcal{S}_i(x_{a \rightarrow b}^k), y_a) + \lambda \mathcal{L}_{Focal}(\mathcal{S}_i(x_{a \rightarrow b}^k), y_a),$$

where $i = 0$ represents the primary prediction branch, $i = \{1, \dots, 5\}$ represents i -th prediction branch, $k = \{1, 2, \dots, n_k\}$ denotes the k -th style to generate the image, and $\lambda = 10$ is the trade-off parameter to control the importance of each loss.

IV. EXPERIMENTS

A. Experimental Setup

1) *Datasets*: In our experiment, two datasets are utilized to evaluate the efficacy of our method. The first dataset is the vestibular schwannoma segmentation dataset [4], which consists of unpaired 105 contrast-enhanced T1 (ceT1) and 105 high-resolution T2 (hrT2) Magnetic Resonance Imaging (MRI) volumes from different clinical sites. We complete adaptation experiments only on the direction of “ceT1 to hrT2” since the annotations of hrT2 volumes are inaccessible, and the evaluation of UDA results are obtained by submitting the results online². The segmentation labels contain two cardiac structures, cochlea and vestibular schwannoma (VS), which are very small targets as shown in Figure 1 (c).

The second dataset includes 30 volumes of CT data collected from [36], and 20 volumes of T2-SPIR MRI training data collected from the ISBI 2019 CHAOS Challenge [37]. To evaluate the performance of UDA, we conduct the experiments both in the “CT to MRI” direction and in the “MRI to CT” direction. There are four abdominal organs to be segmented including the liver, right kidney, left kidney, and spleen.

For vestibular schwannoma datasets, the split of training and test set are given officially. For the multi-organ abdominal datasets, we split the training-test sets according to SIFA [5] for a fair comparison. Since CT data include the area from neck to knee while MRI data only contain the abdominal area, we crop the CT images to have the same view with MRI. In order to meet the requirement of our anisotropic architecture, the two datasets were first spatially normalized to the spacing of [1.5, 0.41, 0.41] and [4, 1, 1] respectively, where the ratio of the through-plane and in-plane resolution is around 4. Then all the volumes are padded to have the same size of 512×512 pixels on XY plane. We perform min-max normalization to rescale the images intensity to the the range $[-1, 1]$ in the data pre-processing stage.

2) *Implementation Details*: We use one RTX 3090 GPU (24G memory) to carry out our experiments. For the training of style transfer, we use LSGANs [38] to stabilize the training. The parameters of disentangled GAN are optimized using

²Official evaluation website: <https://crossmoda.grand-challenge.org/>

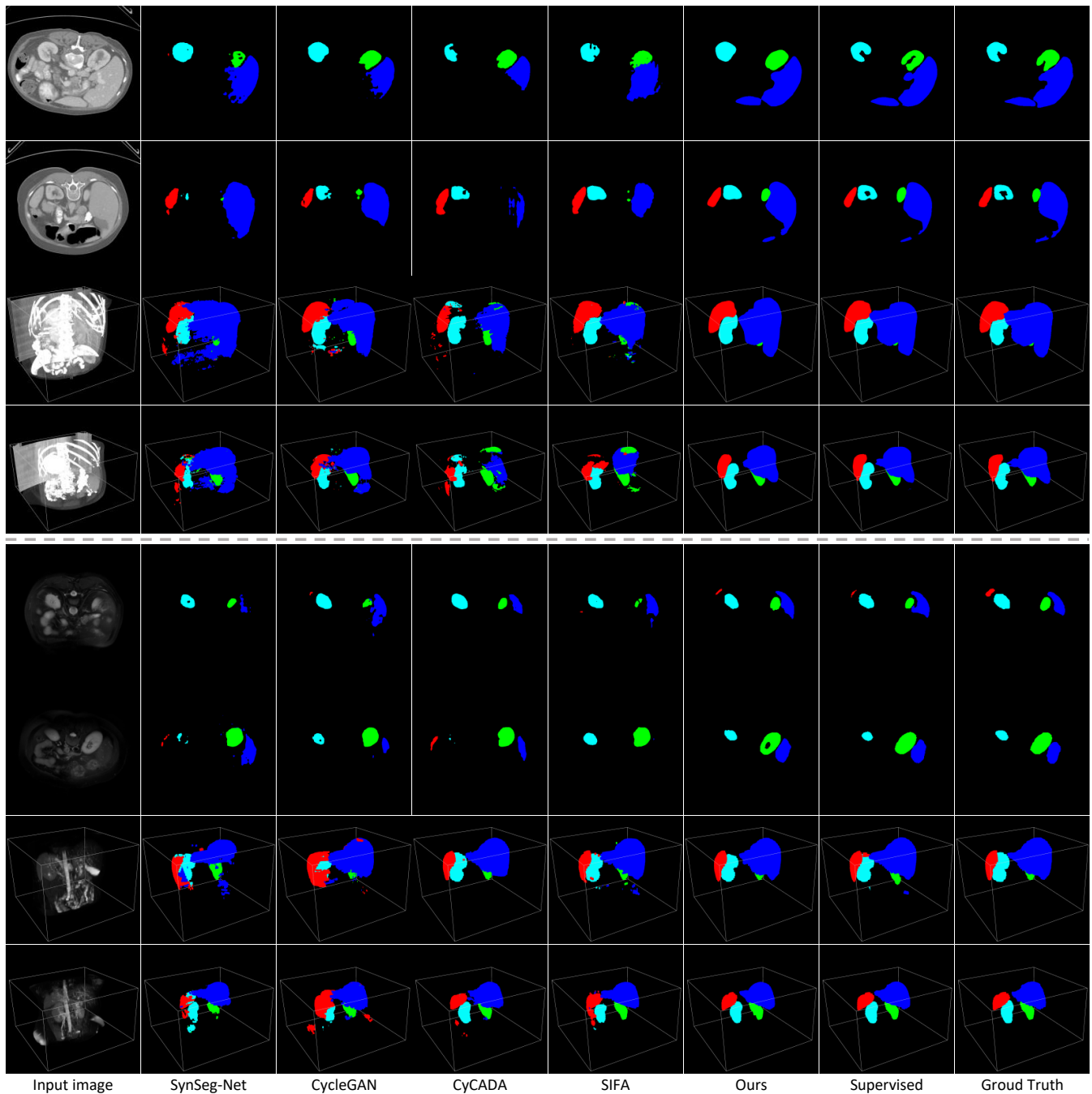


Fig. 6. Visualization of segmentation results produced by different methods for abdominal CT images (top four rows) and MRI images (bottom four rows). 2D visual comparison (1st,2nd,5th,6th rows) for slices and 3D visual comparison (3rd,4th,7th,8th rows) for volumes are provided. From left to right are the raw test images (1st column), results of other 2D unsupervised domain adaptation methods (2nd-5th columns), results of our 3D framework (6th column), results of our proposed DAR-UNet under supervised training (7th column) and ground truth (last column). The liver, right kidney, left kidney and spleen are indicated in blue, green, cyan, and red color respectively. Each row corresponds to one example.

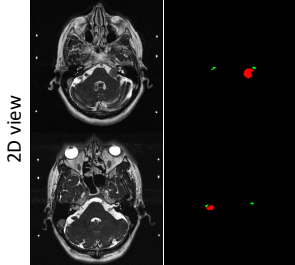
AdaBelief [39] for 50 epochs and updated following the rule of TTUR [40], in which discriminators and generators have the different learning rates of $2e-4$ and $1e-4$. For the training of segmentation, the parameters are optimized using AdaBelief for 100 epochs with an initial learning rate of $5e-4$ and a cosine learning rate decay strategy. We train DAR-UNet with a batch size of 2, and the sub-volumes with the size of $32 \times 256 \times 256$ voxels are randomly cropped during training, followed by data augmentation including random rotation,

random translation, and elastic transform. The training time for style transfer and segmentation are both about 6 hours. We set $n_k = 5$ and $\lambda = 10$ for all the experiments. During inference, we engage a sliding window strategy with size of $32 \times 256 \times 256$ and stride of $16 \times 128 \times 128$, thus we can handle any size of input volumes.

3) Evaluation Metrics: Two evaluation metrics are used in our experiment. Dice similarity coefficient (Dice) measures the voxel-level overlap ratio between the prediction mask and

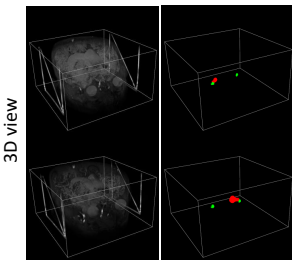
TABLE I
PERFORMANCE COMPARISON OF DIFFERENT METHODS FOR CROSS-MODALITY VESTIBULAR SCHWANNOMA SEGMENTATION.

Vestibular schwannoma ceT1 - hrT2						
Method	Dice(%) \uparrow			ASD(voxel) \downarrow		
	Cochlea	VS	Avg	Cochlea	VS	Avg
No adaptation	0.00	0.00	0.00	-	-	-
SynSeg-Net [13]	3.72	21.99	12.86	11.99	14.68	13.34
CycleGAN [8]	27.69	29.86	28.78	4.65	14.34	9.50
CyCADA [15]	30.58	33.08	31.83	2.32	14.05	8.19
SIFA [5]	39.06	35.73	37.40	1.08	13.34	7.21
Ours	80.22	84.44	82.33	0.20	0.59	0.40



2D view

Input Ours



3D view

Input Ours

(Right) 2D and 3D visualization of segmentation (by our method). Cochlea and vestibular schwannoma are indicated in green & red, respectively.

TABLE II
PERFORMANCE COMPARISON OF DIFFERENT METHODS FOR CROSS-MODALITY ABDOMINAL SEGMENTATION.

Method	AbdominalMRI-CT					AbdominalCT-MRI														
	Dice(%) \uparrow					ASD(voxel) \downarrow					Dice(%) \uparrow					ASD(voxel) \downarrow				
	Liver	R.kid	L.Kid	Spleen	Avg	Liver	R.kid	L.Kid	Spleen	Avg	Liver	R.kid	L.Kid	Spleen	Avg	Liver	R.kid	L.Kid	Spleen	Avg
Supervisedtraining	95.77	89.93	90.00	90.48	91.54	1.18	2.38	0.94	1.51	1.50	94.98	94.13	92.22	93.21	93.64	0.54	0.32	0.50	0.56	0.48
Noadaptation	8.89	1.56	0.00	0.00	2.63	29.91	62.14	68.85	79.10	60.00	0.00	0.00	0.00	0.00	0.00	-	-	-	-	-
SynSeg-Net [13]	85.00	82.10	72.70	81.00	80.20	2.20	1.30	2.10	2.00	1.90	87.20	90.20	76.60	79.60	83.40	2.80	0.70	4.80	2.50	2.70
AdaOutput [14]	85.40	79.70	79.70	81.70	81.60	1.70	1.20	1.80	1.60	1.60	85.80	89.70	76.30	82.20	83.50	1.90	1.40	3.00	1.80	2.10
CycleGAN [8]	83.40	79.30	79.40	77.30	79.90	1.80	1.30	1.20	1.90	1.60	88.80	87.30	76.80	79.40	83.10	2.00	3.20	1.90	2.60	2.40
CyCADA [15]	84.50	78.60	80.30	76.90	80.01	2.60	1.40	1.30	1.90	1.80	88.70	89.30	78.10	80.20	84.10	1.50	1.70	1.30	1.60	1.50
SIFA [5]	88.00	83.30	80.90	82.60	83.70	1.20	1.00	1.50	1.60	1.30	90.00	89.10	80.20	82.30	85.40	1.50	0.60	1.50	2.40	1.50
DSAN [35]	-	-	-	-	-	-	-	-	-	-	89.30	90.16	90.09	89.84	89.84	-	-	-	-	-
Ours	89.59	86.10	88.10	84.38	87.04	1.73	1.42	2.02	2.15	1.83	94.00	90.50	87.99	92.46	91.25	0.64	0.50	0.67	0.38	0.54

the ground truth mask. Average symmetric surface distance (ASD) is used to calculate the average distances between the surface of the two in 3D. Higher Dice value as well as lower ASD indicate better segmentation performance.

B. Comparison with State-of-the-art Methods

We compare our proposed method with the state-of-the-art methods in the vestibular schwannoma and abdominal multi-organ segmentation tasks. Several 2D-based deep learning methods are included, e.g., SynSeg-Net [13], AdaOutput [14], CycleGAN [8], CyCADA [15], SIFA [5], and DSAN [35]. SynSeg-Net and CycleGAN adapt image appearance for pixel alignment, while AdaOutput engages feature alignment for domain adaptation. The rest of them conduct a mixture of image and feature alignments, while none employs multi-style image generation. In addition, we provide the results of our proposed DAR-UNet under supervised training and no adaptation in abdominal datasets, which can be approximately considered as the upper and lower bounds of our method. We do not give the upper bound on the vestibular schwannoma dataset since the target domain annotations are unavailable. The comparative results on vestibular schwannoma segmentation are obtained by reimplementing the official codes, while the results on abdominal multi-organ segmentation are referred to previous works [5], [35]. To our best knowledge, there is no other comparable 3D-based UDA methods that were developed for vestibular schwannoma and abdominal multi-organ datasets.

The quantitative performance of different methods are presented in Table I (Left) and Table II for brain and abdomi-

nal images respectively. Obviously, our method significantly outperforms other comparative approaches by a large margin, especially on the vestibular schwannoma dataset. Classical 2D UDA methods fail to perform well on this task since the background takes up a great proportion ($\geq 99\%$). Most of the 2D slices do not contain any key structures, so the network tends to output the all-black labels due to severe class-imbalance problems. Notably, the methods utilizing both image and feature alignments (e.g., CyCADA, SIFA, and DSAN) obtain better results than those that adopt one single alignment. This result confirms our statement that feature alignment may help the incomplete image alignment that exploits a domain-level style transfer. Meanwhile, even if we do not use feature alignment, our method achieves promising results thanks to the complete image alignment in the first stage. In addition, our 3D segmentor as well as the proposed VAM and QAM can further boost the performance. It is noted that although our method outperforms the other approaches in most cases, the ASD result of the direction MRI-CT on abdominal multi-organ dataset is worse than the other methods. It may be caused by the different criteria of annotation for MRI and CT images. For instance, the annotation of kidney is hollow in MRI but filled in CT, as shown in Figure 6 last column.

The visualization of different methods on abdominal images is presented in Figure 6. Since the previous methods took the 2D framework, the interactions across depth are missed, resulting in some inconsistent results along Z-axis. However, this inconsistency cannot be observed in the 2D view, we also provided 3D visualization of the results. Our method can fully

TABLE III
PERFORMANCE COMPARISON OF DAR-UNET TRAINED USING
TRANSFERRED IMAGES FROM DIFFERENT GANS.

	MRI - CT Dice (%)	CT - MRI Dice (%)
Entangled GAN	81.37	84.87
Disentangled GAN (mean-style generation)	83.87	89.84
Disentangled GAN (multi-style generation)	87.04	91.25

take advantage of 3D framework to model depth information as well as both VAM and QAM to capture global features, thus leading to more consistent results along Z -axis.

C. Effect of Complete Image Alignment

The main goal of image alignment is to reduce the domain gap from a stylistic perspective, weakening the negative influence of different appearances between two domains on the segmentation task. Therefore, we design the network according to the following criteria: 1) the semantic content of the domain should be well preserved, and 2) the distribution of transferred source images and target images should fit well. To achieve this, we engage a feature disentangled GAN to extract the domain-invariant representations as content and the domain-specific representations as style. Once trained, we can sample the styles from target domain images for each content extracted from source domain samples, thus making the transferred samples own the target domain distribution.

To evaluate the effect of complete image alignment, we conduct several experiments on abdominal multi-organ data with different settings of style transfer. Specifically, we train our DAR-UNet using the transferred images obtained from entangled GAN, disentangled GAN with the mean-style generation, and disentangled GAN with multi-style generation. We utilize CycleGAN as the entangled GAN considering its popular usage in computer vision. As shown in Table III, the entangled GAN attains only 81.37% and 84.87% in Dice in two directions, the lowest among the three methods. Although the disentangled way improves both the tasks, it is still lower than the representation disentangled multi-style generation synthesis by 3.17% and 1.41% in MRI-CT and CT-MRI respectively, demonstrating the effectiveness of multi-style transformation in reducing domain gap. We also visualize our disentangled GAN in Figure 7. As observed, the generated images vary from each other; this is especially evident in some samples. For instance, the style 2 samples in ceT1 - hrT2 and the style 4 samples in MRI - CT exhibit quite different appearances from the mean style samples. Namely, if only mean style samples are used to train the segmentation net, samples with rare styles in the target domain may not be correctly segmented as the network has never seen this kind of distribution before. This phenomenon is also revealed in the numerical result. We can infer that the style variance in the CT domain is much larger than MR, which means the multi-style transformation strategy should be more useful in task MRI-CT segmentation. If we compare the improvement brought by multi-style generation, it is evident that this value is higher for the MRI-CT adaptation task than its reverse direction.

TABLE IV
ABLATION STUDY ON THE KEY COMPONENTS OF DAR-UNET.

Anisotropic stride	-	✓	✓	✓	✓	✓
QAM	-	-	✓	-	✓	✓
VAM	-	-	-	✓	✓	✓
Deep Supervision	-	-	-	-	-	✓
Dice (%)	78.60	86.84	88.81	88.86	89.96	91.25
Memory Cost During Training (GB)	43.6	14.6	18.4	17.0	21.2	23.5

D. Ablation Study

To examine different components in our framework, we conduct an ablation study with five variants of the proposed DAR-UNet on the abdominal dataset for CT to MRI adaptation, as shown in Table IV. First, we remove all the components to get the traditional 3D ResU-Net with isotropic resolution. We train the model using the spatial normalized patches with a size of (128, 256, 256) and batch size of 2, taking 43.6 GB memory in total. Second, after adding the proposed anisotropic architecture and training the model with anisotropic patches with a size of (32, 256, 256), the memory cost dramatically decreases to 14.6 GB and the performance increases by 8.24%. The intuition behind the anisotropic stride strategy is to increase the information abundance of volume while reducing useless features without dropping the learnable parameters. Next, the proposed QAM and VAM improves 1.97% and 2.02% in Dice score respectively suggesting that attention modules modeling global correlation can significantly benefit 3D segmentation tasks. Meanwhile, although the improvement brought by the attention module is close to saturation, the combination of QAM and VAM achieved almost 1% improvement compared with using either one of them. Finally, a deep supervision strategy further boosts the performance of 1.29% by alleviating the gradient vanishing problem in the deep level of the DAR-UNet, with an additional memory cost of 2.3 GB only.

V. DISCUSSION AND LIMITATION

In this paper, we carefully design the proposed VAM and QAM for maximizing performance improvement under certain memory costs. To be specific, VAM is designed to replace the skip connection in UNet-like architecture, and then enhance the encoder-guided feature map; QAM is designed to capture the cross-dependencies among channel and spatial dimensions in the 3D task. The two different attention modules are designed to build interdependencies among channels or/and spatial locations. With the robust and strong segmentor, the proposed DAR-UNet can be trained on target-like source datasets, and generalize well on target domain.

Meanwhile, although we handle the issue of UDA problem with a novel 3D framework, our method still has some limitations. First, since a 3D image-to-image translation framework may require tons of training samples and GPU memory, we engaged a 2D image-to-image translation by transferring the slices of volumes for efficiency, which may result in inconsistent style along axial. Second, as a two-step framework,

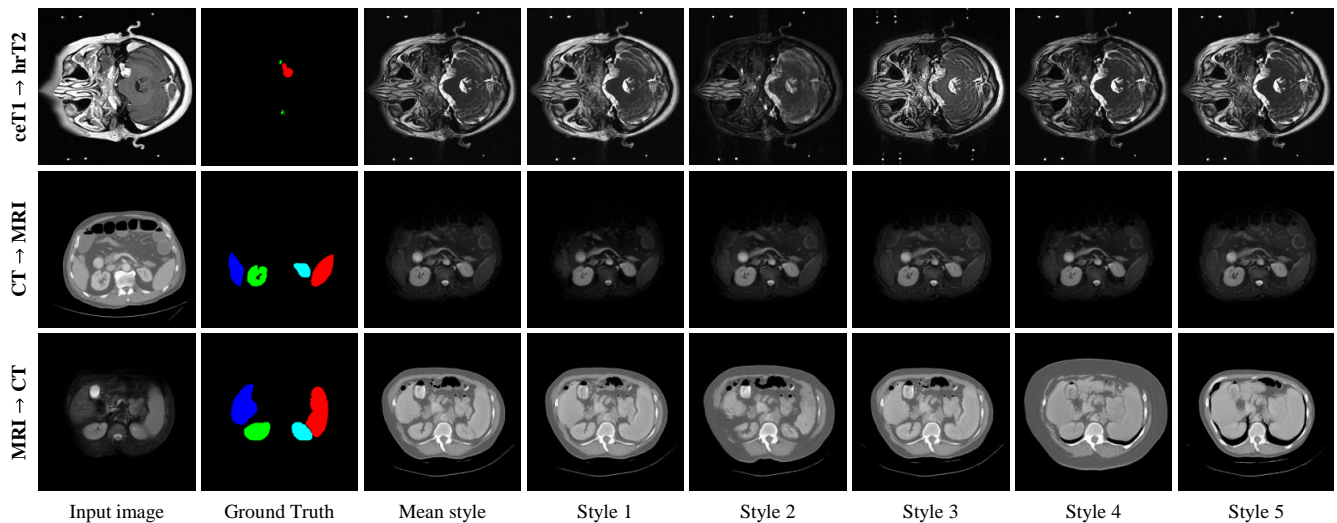


Fig. 7. Visual comparison of 2D image transfer results produced by mean-style transformation and multi-style transformation for the direction of 'ceT1-hrT2' (1st row), 'CT-MRI' (2nd row) and 'MRI-CT' (3rd row). From left to right are the raw test images (1st column), ground truth (2nd column), results of partial image alignment (3rd column) and results of complete image alignment (4-8th columns). Style 1-5 indicate different sampled styles using K-Means.

the feature alignment, and semantic segmentation cannot be optimized simultaneously, which may lower the performance ceiling of the whole model. Future work includes investigation of intra-domain variance in medical datasets and exploration of a fully end-to-end 3D unsupervised domain adaptation framework.

VI. CONCLUSION

In this paper, we propose a novel 3D framework for unsupervised domain adaptation in medical image segmentation. Combining multi-style transformation and dual-attention modules, our framework shows impressive performance in alleviating domain shift problems. In two cross-modality tasks, vestibular schwannoma and multi-organ abdominal segmentation, our proposed approach greatly exceeds the performance of state-of-the-art methods.

ACKNOWLEDGEMENT

The work was partially supported by the following: National Natural Science Foundation of China under no. 61876155; Jiangsu Science and Technology Programme (Natural Science Foundation of Jiangsu Province) under no. BE2020006-4; Key Program Special Fund in XJTLU under no. KSF-T-06 and no. KSF-E-37; Research Development Fund in XJTLU under no. RDF-19-01-21.

REFERENCES

- [1] A. S. Panayides, A. Amini, N. D. Filipovic, A. Sharma, S. A. Tsiftaris, A. Young, D. Foran, N. Do, S. Golemati, T. Kurc *et al.*, "Ai in medical imaging informatics: current challenges and future directions," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 7, pp. 1837–1857, 2020.
- [2] K. Yao, K. Huang, J. Sun, L. Jing, D. Huang, and C. Jude, "Scaffold-a549: a benchmark 3d fluorescence image dataset for unsupervised nuclei segmentation," *Cognitive Computation*, vol. 13, no. 6, pp. 1603–1608, 2021.
- [3] M. Mahmud, M. S. Kaiser, T. M. McGinnity, and A. Hussain, "Deep learning in mining biological data," *Cognitive computation*, vol. 13, no. 1, pp. 1–33, 2021.
- [4] J. Shapey, A. Kujawa, R. Dorent, G. Wang, A. Dimitriadis, D. Grishchuk, I. Paddick, N. Kitchen, R. Bradford, S. R. Saeed *et al.*, "Segmentation of vestibular schwannoma from mri, an open annotated dataset and baseline algorithm," *Scientific Data*, vol. 8, no. 1, pp. 1–6, 2021.
- [5] C. Chen, Q. Dou, H. Chen, J. Qin, and P. A. Heng, "Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 39, no. 7, pp. 2494–2505, 2020.
- [6] K. Huang, A. Hussain, Q.-F. Wang, and R. Zhang, *Deep learning: fundamentals, theory and applications*. Springer, 2019, vol. 2.
- [7] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision*. IEEE, 2016, pp. 565–571.
- [8] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2223–2232.
- [9] K. Yao, J. Sun, K. Huang, L. Jing, H. Liu, D. Huang, and C. Jude, "Analyzing cell-scaffold interaction through unsupervised 3d nuclei segmentation," *International journal of bioprinting*, vol. 8, no. 1, 2022.
- [10] Q. Dou, C. Ouyang, C. Chen, H. Chen, B. Glocker, X. Zhuang, and P.-A. Heng, "Pnp-adanet: Plug-and-play adversarial domain adaptation network with a benchmark at cross-modality cardiac segmentation," *IEEE Access*, vol. 7, pp. 99 065–99 076, 2019.
- [11] Z. Gao, S. Zhang, K. Huang, Q. Wang, and C. Zhong, "Gradient distribution alignment certificates better adversarial domain adaptation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8937–8946.
- [12] S. Ben-David, J. Blitzer, K. Crammer, F. Pereira *et al.*, "Analysis of representations for domain adaptation," *Advances in Neural Information Processing Systems*, vol. 19, p. 137, 2007.
- [13] Y. Huo, Z. Xu, H. Moon, S. Bao, A. Assad, T. K. Moyo, M. R. Savona, R. G. Abramson, and B. A. Landman, "Synseg-net: Synthetic segmentation without target modality ground truth," *IEEE Transactions on Medical Imaging*, vol. 38, no. 4, pp. 1016–1025, 2018.
- [14] Y.-H. Tsai, W.-C. Hung, S. Schultze, K. Sohn, M.-H. Yang, and M. Chandraker, "Learning to adapt structured output space for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7472–7481.
- [15] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell, "Cycada: Cycle-consistent adversarial domain adapta-

- tion,” in *International Conference on Machine Learning*. PMLR, 2018, pp. 1989–1998.
- [16] H. Kim and A. Mnih, “Disentangling by factorising,” in *International Conference on Machine Learning*. PMLR, 2018, pp. 2649–2658.
- [17] Y.-C. Liu, Y.-Y. Yeh, T.-C. Fu, S.-D. Wang, W.-C. Chiu, and Y.-C. F. Wang, “Detach and adapt: Learning cross-domain disentangled deep representation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8867–8876.
- [18] A. H. Liu, Y.-C. Liu, Y.-Y. Yeh, and Y.-C. F. Wang, “A unified feature disentangler for multi-domain image translation and manipulation,” *Conference on Neural Information Processing Systems*, 2018.
- [19] H.-Y. Lee, H.-Y. Tseng, Q. Mao, J.-B. Huang, Y.-D. Lu, M. Singh, and M.-H. Yang, “Drit++: Diverse image-to-image translation via disentangled representations,” *International Journal of Computer Vision*, vol. 128, no. 10, pp. 2402–2417, 2020.
- [20] L. Van der Maaten and G. Hinton, “Visualizing data using t-sne,” *Journal of Machine Learning Research*, vol. 9, no. 11, 2008.
- [21] J. A. Hartigan and M. A. Wong, “Algorithm as 136: A k-means clustering algorithm,” *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979.
- [22] K. Kawakami, “Supervised sequence labelling with recurrent neural networks,” *Ph. D. thesis*, 2008.
- [23] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [24] Q. Yan, B. Wang, W. Zhang, C. Luo, W. Xu, Z. Xu, Y. Zhang, Q. Shi, L. Zhang, and Z. You, “An attention-guided deep neural network with multi-scale feature fusion for liver vessel segmentation,” *IEEE Journal of Biomedical and Health Informatics*, 2020.
- [25] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, “Residual attention network for image classification,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3156–3164.
- [26] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, “Self-attention generative adversarial networks,” in *International Conference on Machine Learning*. PMLR, 2019, pp. 7354–7363.
- [27] D. Misra, T. Nalamada, A. U. Arasanipalai, and Q. Hou, “Rotate to attend: Convolutional triplet attention module,” in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2021, pp. 3139–3148.
- [28] X. Huang and S. Belongie, “Arbitrary style transfer in real-time with adaptive instance normalization,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1501–1510.
- [29] Z. Zhang, Q. Liu, and Y. Wang, “Road extraction by deep residual u-net,” *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 749–753, 2018.
- [30] D. Jha, P. H. Smedsrud, M. A. Riegler, D. Johansen, T. De Lange, P. Halvorsen, and H. D. Johansen, “Resunet++: An advanced architecture for medical image segmentation,” in *2019 IEEE International Symposium on Multimedia*. IEEE, 2019, pp. 225–2255.
- [31] D. Misra, T. Nalamada, A. U. Arasanipalai, and Q. Hou, “Rotate to attend: Convolutional triplet attention module,” in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2021, pp. 3139–3148.
- [32] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, “Deeply-supervised nets,” in *Artificial Intelligence and Statistics*. PMLR, 2015, pp. 562–570.
- [33] F. Milletari, N. Navab, and S.-A. Ahmadi, “V-net: Fully convolutional neural networks for volumetric medical image segmentation,” in *2016 Fourth International Conference on 3D Vision*. IEEE, 2016, pp. 565–571.
- [34] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in *Proceedings of IEEE International Conference on Computer Vision*, 2017, pp. 2980–2988.
- [35] X. Han, L. Qi, Q. Yu, Z. Zhou, Y. Zheng, Y. Shi, and Y. Gao, “Deep symmetric adaptation network for cross-modality medical image segmentation,” *IEEE transactions on medical imaging*, vol. 41, no. 1, pp. 121–132, 2022.
- [36] B. Landman, Z. Xu, J. E. Iglesias, M. Styner, and a. A. K. T. R. Langerak, “Multi-atlas labeling beyond the cranial vault,” <https://www.synapse.org/Synapse:syn3193805/wiki/217789>, 2017.
- [37] A. E. Kavur, N. S. Gezer, M. Barış, S. Aslan, P.-H. Conze, V. Groza, D. D. Pham, S. Chatterjee, P. Ernst, S. Özkan *et al.*, “Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation,” *Medical Image Analysis*, vol. 69, p. 101950, 2021.
- [38] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, “Least squares generative adversarial networks,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2794–2802.
- [39] J. Zhuang, T. Tang, Y. Ding, S. Tatikonda, N. Dvornek, X. Papademetris, and J. Duncan, “Adabelief optimizer: Adapting stepsizes by the belief in observed gradients,” *Conference on Neural Information Processing Systems*, 2020.
- [40] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “Gans trained by a two time-scale update rule converge to a local nash equilibrium,” *Advances in Neural Information Processing Systems*, vol. 30, 2017.