DAR-UNet: Dual Attention ResU-Net for CrossMoDa Challenge

Kai Yao[1,2], Zixian Su[1,2], Xi Yang[1], Kaizhu Huang[1], and Jie Sun[1]

[1] School of Advanced Technology, Xi'an Jiaotong-Liverpool University, China
[2] University of Liverpool, the United Kingdom

Abstract. Unsupervised domain adaptation(UDA) methods have shown promising performance in the cross-domain tasks, which aim to alleviate the retraining problem on target domain without annotated labels. However, there are still huge challenges to do this in medical imaging due to the severe domain shift between the source domain and target domain. In this paper, we propose a novel framework composed of Content-style disentangled image transfer and Dual Attention ResU-Net to better reduce the domain shift, which proves effective and intuitive to transfer the learned knowledge from the source domain to the target domain in medical image segmentation tasks.

1 Introduction

Contrast-enhanced T1 (ceT1) Magnetic Resonance Imaging (MRI) scans are commonly used for vestibular schwannoma (VS) segmentation, while ceT1 requires patients to be injected intravenously with a contrast agent to achieve its signal-enhancing effect, which may cause allergies in some people and is very costly. Recent work has demonstrated that high-resolution T2 (hrT2) imaging could be a reliable, safer, and lower-cost alternative to ceT1 [1,2,3].

For these reasons, we propose an unsupervised domain adaptation framework for cross-modality task (from ceT1 to hrT2) that can automatically perform VS and cochlea segmentation on hrT2 scans. The proposed method can realize both medical volume translation as well as semantic segmentation. Specifically, a generator is first learned for cross-domain volume-to-volume translation, and then a segmentor is trained taking advantage of the translated images from the first stage. Different from the previous efforts which encode domains into a common feature space, our method extracts the domain-invariant features and domain-specific features separately and specifically formulate the domain-specific features into the framework of UDA problems. This framework facilitates the encoder to express the shared and specific information separately and achieve domain adaptation both on image level and feature level. Inspired by the convolutional Triplet Attention Module which can capture cross-dimension between the spatial dimensions and channel dimensions (C, H, W) of the input, a convolutional Quartet Attention Module is adopted in segmentation part that also stresses the adjacent semantic information between slices on higher dimensions(N, C, H, W). Combined with the DAR-UNet as a segmentation framework, our model can achieve superior result in experiment.

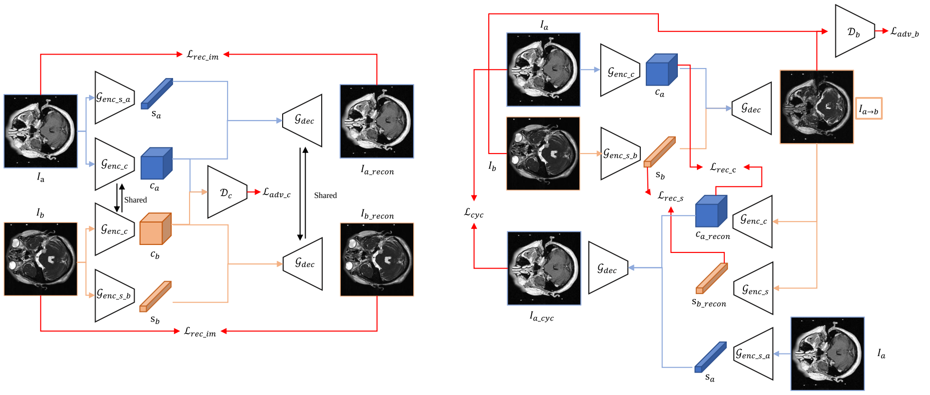

Fig. 1. Training scheme of C-S disentangled image-to-image translation.

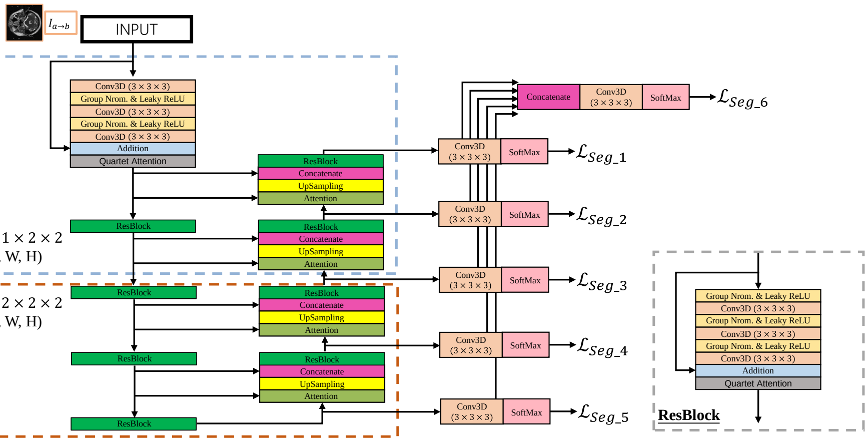

Fig. 2. Architecture of our DAR-UNET.

2 Method

This section will first provide an overview of our proposed framework for unsupervised domain adaptation in task cross-modality medical image segmentation in Sec. 2.1. The objective functions and overall loss will then be given in Sec. 2.2. Finally, we will discuss the training details in Sec. 2.3.

2.1 Method Overview

In this work, we focus on the problem of unsupervised domain adaptation for medical image semantic segmentation, where we have access to the labeled source dataset $\{x_a, y_a\}$ and unlabeled target set $\{x_b\}$. The overall framework is shown in Fig. 1. It is composed of a synthesis subnet and a segmentation subnet which are introduced as follows.

Feature-Disentanglement-Based Synthesis Subnet As shown in Fig. 1, we use a 2D auto-encoder architecture as the synthesis subnet. Similar to the MUNIT [4], we first assume that the latent space of images can be decomposed into content space and style space. Following this assumption, we disentangle features of each domain into their domain-invariant features (DIFs) and domain-specific features (DSFs) using a shared content encoder $\mathcal{G}_{enc\_c}$ and domain-specific style encoder $\mathcal{G}_{enc\_s\_a}, \mathcal{G}_{enc\_s\_b}$. Then, a content discriminator $\mathcal{D}_c$ is adopted to distinguish the content c_a, c_b extracted by $\mathcal{G}_{enc\_c}$ in order to align the content representations of the source and target domain. Simultaneously, the content representation c_a and the style representation s_a can be composed into reconstructed image I_{a_recon} by the decoder $\mathcal{G}_{dec}$, i.e., $I_{a\_recon} = \mathcal{G}_{dec}(\mathcal{G}_{enc\_c}(I_a), \mathcal{G}_{enc\_s\_a}(I_a))$. I_b goes by the same expressions where I_a and I_b denote input images from different domains. The Image-to-image translation is performed by swapping the content and the style extractions. In our training progress, we randomly sample the style code of one domain and recombine it with the content code to produce fake images, i.e., $I_{a\to b} = \mathcal{G}_{dec}(\mathcal{G}_{enc\_c}(I_a), \mathcal{G}_{enc\_s\_b}(I_b))$. In addition, we follow the existing works [4,5] to adopt the multi-scale discriminator as the image discriminator $\mathcal{D}_a, \mathcal{D}_b$ for both domains, which learns to determine whether an image is a real image of its domain or a fake one generated by $\mathcal{G}$.

Segmentation Subnet via DAR-UNet For segmentation, we utilize DAR-UNet as the backbone structure to predict the synthesized images $I_{a\to b}$, as described in Fig. 2. The DAR-UNet is adapted from the classic 2D ResU-Net [6,7] to extend to volumetric data by replacing part of the 2D convolutions with 3D convolutions, providing accurate segmentation masks for the 3D dataset be computationally efficient. Inspired by previous works [8,9], we keep the depth unchanged in the first and second stages of Unet. Additionally, we apply a convolutional quartet attention module inspired by the previous work, triplet attention [10], that captures dependencies between the (C, H, W), (C, W, N), (H, W, N) and (C, H, N) dimensions of the input tensor respectively. This quartet attention module works in a four-branched 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 $\mathcal{X} \in R^{N\times C\times H\times W}$, we first pass it to each of the four branches in the proposed quartet attention module. In the first branch, we build interactions between number, height and width (N, H, D) dimension. To achieve so, the input $\mathcal{X}$ first rotated 90 degrees to become $\hat{\mathcal{X}_1}$, where $\hat{\mathcal{X}_1}$ is of shape $(C \times N \times H \times W)$. After that, $\hat{\mathcal{X}_1}$ is passed through Z-pool, which is for reducing the zeroth dimension of the tensor to two by concatenating the average pooled and max pooled features across that dimension, and $\hat{\mathcal{X}_1}$ is subsequently reduced to $\hat{\mathcal{X}_1^*}$ with shape $(2 \times N \times H \times W)$. $\hat{\mathcal{X}_1^*}$ is then passed through a standard convolutional layer of kernel size $k \times k \times k$ followed by a batch normalization layer, which provides the intermediate output of dimensions $(1 \times N \times H \times W)$. The resultant attention weights are generated by passing the tensor through a sigmoid activation layer σ, and subsequently applied to $\hat{\mathcal{X}_1^*}$. This process is similar for the remaining branches. To summarize, the whole process to obtain the refined attention-applied tensor y from the quatert attention can be decribed by the following equation:

$$y = \frac{1}{4}(\overline{\hat{\mathcal{X}}_1\sigma(\psi_1(\hat{\mathcal{X}}_1^*))} + \overline{\hat{\mathcal{X}}_2\sigma(\psi_2(\hat{\mathcal{X}}_2^*))} + \overline{\hat{\mathcal{X}}_3\sigma(\psi_3(\hat{\mathcal{X}}_3^*))} + \mathcal{X}\sigma(\psi_4(\hat{\mathcal{X}}_4))), \tag{1}$$

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.

2.2 Objective Functions

In the first stage, the training loss consists of reconstruction loss, adversarial loss and cycle consistency loss. In order to build one to one mapping among image space, style space and content space, the reconstruction loss is performed:

$$\begin{aligned}
&\mathcal{L}_{rec}(\mathcal{G}_{enc\_c}, \mathcal{G}_{enc\_s\_a}, \mathcal{G}_{enc\_s\_b}, \mathcal{G}_{dec}) \\
&\quad = \lambda_{rec\_im}(\|I_a - I_{a\_recon}\|_1 + \|I_b - I_{b\_recon}\|_1) \\
&\qquad + \lambda_{rec\_c}(\|c_a - c_{a\_recon}\|_1 + \|c_b - c_{b\_recon}\|_1) \\
&\qquad + \lambda_{rec\_s}(\|s_a - s_{a\_recon}\|_1 + \|s_b - s_{b\_recon}\|_1),
\end{aligned} \tag{2}$$

where I denotes the 2D slices of x, c denotes the disentangled content, s denotes the disentangled style, λ_{rec_im}, λ_{rec_c}, λ_{rec_s} equal to $10, 1, 1$, respectively.

The adversarial loss is composed of feature-level and image-level losses, which is shown as following:

$$\begin{aligned}
&\mathcal{L}_{adv}(\mathcal{G}_{enc\_c}, \mathcal{G}_{enc\_s\_a}, \mathcal{G}_{enc\_s\_b}, \mathcal{G}_{dec}, \mathcal{D}_a, \mathcal{D}_b, \mathcal{D}_c) \\
&\quad = (\mathcal{D}_a(I_{b\to a}) - 0)^2 + (\mathcal{D}_a(I_a) - 1)^2 + (\mathcal{D}_b(I_{a\to b}) - 0)^2 \\
&\qquad + (\mathcal{D}_b(I_b) - 1)^2 + (\mathcal{D}_c(c_a) - 0)^2 + (\mathcal{D}_c(c_b) - 1)^2.
\end{aligned} \tag{3}$$

The cycle consistency loss is shown as following:

$$\mathcal{L}_{cyc}(\mathcal{G}_{enc\_c}, \mathcal{G}_{enc\_s\_a}, \mathcal{G}_{enc\_s\_b}, \mathcal{G}_{dec}) = \lambda_{cyc}(\|I_a - I_{a\_cyc}\|_1 + \|I_b - I_{b\_cyc}\|_1), \tag{4}$$

where $\lambda_{cyc} = 10$. Finally, we jointly train the encoder, decoder, and discriminators to optimize the full objective in the first stage:

$$\underset{\mathcal{G}_{enc\_c}, \mathcal{G}_{enc\_s\_a}, \mathcal{G}_{enc\_s\_b}, \mathcal{G}_{dec}}{\arg\min} \; \underset{\mathcal{D}_a, \mathcal{D}_b, \mathcal{D}_c}{\arg\max} \; \mathcal{L}_{adv} + \mathcal{L}_{rec} + \mathcal{L}_{cyc}.$$

In the second stage, a deep supervision strategy is used in our framework and the loss for each branch is shown below:

$$\mathcal{L}_{seg} = \mathbb{E}_{x_a, y_a \sim p(X_a, Y_a)}(\ell_{Focal}(x_{a\to b}, y_a) + \ell_{Dice}(x_{a\to b}, y_a)), \tag{5}$$

where ℓ_{Focal} represents the Focal loss [11] and ℓ_{Dice} counts for Dice loss [12]. $x_{a\to b}$ is the synthesis image with the content of the source domain and the style of the target domain.

2.3 Implementation Details

We use one RTX 3090 GPU (24G memory) to carry out our experiment. The images were first normalized to the range $[-1, 1]$, and then random crop, rotation and other augmentation operations in the data reprocessing stage. For the training of stage one, the parameters are optimized using AdaBelief [13] for 50 epochs and updated following the rule of TTUR [14], in which discriminators and generators have the different learning rates of $2e-4$ and $1e-4$. For the training of stage two, the parameters are optimized using AdaBelief for 100 epochs with an initial learning rate of $1e-3$ and a cosine decay strategy is used. We will then describe the detailed composition of the framework.

3 Experimental results

VS DICE	VS ASSD	Cochlea Dice	cochlea ASSD	Mean DICE
0.8423 ± 0.0594	0.6029 ± 0.4073	0.8044 ±0.0428	0.1852 ±0.0586	0.8234 ± 0.0379

Table 1. The result of our method on the validation set.

The results of our method on validation set can be seen in Table 1. Our method obtain promising DICE on VS, while a related low DICE on cochlea. The reason why our method cannot achieve satisfactory DICE on cochlea is that the performance of segment cochlea highly depend on the performance of style transfer on the first stage one. However, a content-style disentangled style transfer can only optimize the global performance of style transfer, resulting in a poor transferred cochlea which is considered a local feature.

4 Conclusion

In this paper, we present a solution to domain adaptation for 3D medical image segmentation by combing the strategy of feature-disentagling style-transfer and 4D-attention-based ResU-Net. The experimental result shows that our model makes superior achievements on the validation set compared with other methods.

References

1. J. Shapey, A. Kujawa, R. Dorent, G. Wang, S. Bisdas, A. Dimitriadis, D. Grishchuk, I. Paddick, N. Kitchen, R. Bradford, S. R. Saeed, S. Ourselin, Segmentation of vestibular schwannoma from magnetic resonance imaging: An open annotated dataset and baseline algorithm, medRxivdoi:10.7937/TCIA.9YTJ-5Q73.
2. J. Shapey, A. Kujawa, R. Dorent, G. Wang, A. Dimitriadis, D. Grishchuk, I. Paddick, N. Kitchen, R. Bradford, S. R. Saeed, S. Bisdas, S. Ourselin, T. Vercauteren, Segmentation of vestibular schwannoma from mri — an open annotated dataset and baseline algorithm, medRxivdoi:10.1101/2021.08.04.21261588.

3. J. Shapey, A. Kujawa, R. Dorent, G. Wang, A. Dimitriadis, D. Grishchuk, I. Paddick, N. Kitchen, R. Bradford, S. R. Saeed, S. Bisdas, S. Ourselin, T. Vercauteren, Segmentation of vestibular schwannoma from mri — an open annotated dataset and baseline algorithm, Scientific DataIn press. Preprint available at medRXiv:10.1101/2021.08.04.21261588.
4. X. Huang, M.-Y. Liu, S. Belongie, J. Kautz, Multimodal unsupervised image-to-image translation, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 172–189.
5. H.-Y. Lee, H.-Y. Tseng, Q. Mao, J.-B. Huang, Y.-D. Lu, M. Singh, M.-H. Yang, Drit++: Diverse image-to-image translation via disentangled representations, International Journal of Computer Vision 128 (10) (2020) 2402–2417.
6. Z. Zhang, Q. Liu, Y. Wang, Road extraction by deep residual u-net, IEEE Geoscience and Remote Sensing Letters 15 (5) (2018) 749–753.
7. D. Jha, P. H. Smedsrud, M. A. Riegler, D. Johansen, T. De Lange, P. Halvorsen, H. D. Johansen, Resunet++: An advanced architecture for medical image segmentation, in: 2019 IEEE International Symposium on Multimedia (ISM), IEEE, 2019, pp. 225–2255.
8. G. Wang, J. Shapey, W. Li, R. Dorent, A. Demitriadis, S. Bisdas, I. Paddick, R. Bradford, S. Zhang, S. Ourselin, et al., Automatic segmentation of vestibular schwannoma from t2-weighted mri by deep spatial attention with hardness-weighted loss, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2019, pp. 264–272.
9. J. Shapey, G. Wang, R. Dorent, A. Dimitriadis, W. Li, I. Paddick, N. Kitchen, S. Bisdas, S. R. Saeed, S. Ourselin, et al., An artificial intelligence framework for automatic segmentation and volumetry of vestibular schwannomas from contrast-enhanced t1-weighted and high-resolution t2-weighted mri, Journal of neurosurgery 134 (1) (2019) 171–179.
10. D. Misra, T. Nalamada, A. U. Arasanipalai, Q. Hou, Rotate to attend: Convolutional triplet attention module, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 3139–3148.
11. T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988.
12. F. Milletari, N. Navab, S.-A. Ahmadi, V-net: Fully convolutional neural networks for volumetric medical image segmentation, in: 2016 fourth international conference on 3D vision (3DV), IEEE, 2016, pp. 565–571.
13. J. Zhuang, T. Tang, Y. Ding, S. Tatikonda, N. Dvornek, X. Papademetris, J. S. Duncan, Adabelief optimizer: Adapting stepsizes by the belief in observed gradients, arXiv preprint arXiv:2010.07468.
14. M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, Gans trained by a two time-scale update rule converge to a local nash equilibrium, Advances in neural information processing systems 30.