An Efficient Cross-Modal Segmentation Method for Vestibular Schwannoma and Cochlea on MRI Images

Cancan Chen[1], Dawei Wang[2], and Rongguo Zhang[3(✉)]

[1] School of Computer Engineering, Jiangsu Ocean University, Lianyungang, China
[2] Infervision Advanced Research Institute, Beijing, China
[3] Academy for Multidisciplinary Studies, Capital Normal University, Beijing, China
zrongguo@cnu.edu.cn

Abstract. To obtain the segmentation results of vestibular schwannoma (VS) and cochlea on high-resolution T2 (hrT2) MR images according to the annotated contrast-enhanced T1 (ceT1) MR images, we propose an efficient cross-modal segmentation framework in this study. An image-to-image model is first applied to transfer ceT1 scans to hrT2 modality to alleviate the domain shift between them. In the model training phase, we adopted data augmentation for both original images and segmentation target regions to adapt to the diversity and heterogeneity of multi-center imaging data. Furthermore, random cropping along the z-axis and random flipping at all axial directions representing the different observation perspectives were implemented. We also utilized a 2.5D ResUnet model as the segmentation backbone. These strategies collectively contribute to improved segmentation output. Eventually, a post-processing method based on image contrast is applied to improve the quality of the pseudo-labels on the hrT2 modality. Our experimental results address the effectiveness of the proposed framework for the crossMoDA23 segmentation task of the vestibular schwannoma and cochlea on hrT2 modality, with average Dice scores of 0.8358 and 0.828 for VS and average Dice scores of 0.8355 and 0.844 for cochlea, respectively on validation and test data.

Keywords: Unsupervised domain adaptation · Cross-modality · Image translation · Heterogeneity · Magnetic resonance imaging (MRI)

1 Introduction

Vestibular schwannoma (VS) is the most common tumor of the cerebellar angular cisterna and the internal auditory canal, and contrast-enhanced T1 (ceT1) MR images are commonly used for the diagnosis and surveillance of patients with VS. Given the risk of gadolinium-containing contrast agents used in scanning ceT1 MR images [1], high-resolution T2 (hrT2) MR image has raised the interest of researchers as a replacement modality given its lower risk and more efficient cost. However, gadolinium-enhanced MR images are required for certain scenarios and

U. Baid et al. (Eds.): crossMoDA 2023/BraTS 2023, LNCS 14669, pp. 355–363, 2024.
https://doi.org/10.1007/978-3-031-76163-8_32

equivocal findings. The goal of CrossMoDA challenge 2021–2023 [4,6,13,15] is to segment VS and cochlea on contrast-enhanced T1(ceT1) and high-resolution T2 (hrT2) MR images for monitoring tumor growth in consecutive follow-ups and even for treatment planning. In this challenge, the aim is to obtain the segmentation results of VS and cochlea on hrT2 MR images with the ceT1 MR images as the gold standard. Besides, the additional segmentation task on the intra- and extra-meatal regions of the tumor from the multi-center heterogeneous scans makes the work more challenging and clinically significant.

Based on the previous excellent work and studies [3,4,14], we propose a simple and efficient cross-modal segmentation framework to tackle the unsupervised domain adaptation segmentation task on unseen data. Firstly, the labeled images on the ceT1 sequence need to be transferred to the hrT2 domain by the image-to-image translation model, such as CycleGAN or CUT [4,10,16]. In this work, we apply CUT as the image-to-image translation model. Secondly, the initial segmentation model for hrT2 sequence is trained on the translated pseudo-hrT2 images and the gold standard annotations on ceT1 sequence, and the pseudo-labels for training target data are obtained. Furthermore, the segmentation model is updated on real hrT2 scans and pseudo-labels by self-training, which could find a better decision boundary on the hrT2 domain [3]. The better experiments on the validation set demonstrate that our approach is effective.

The main contributions of this paper are summarized as follows:

- We propose an efficient cross-modal segmentation method, which could effectively and efficiently perform the segmentation of VS and Cochlea on cross-modality MRI images.
- We first calculate the spacing resolution distribution of all MRI images, especially the thickness on the z-axis, which sparks the helpful ideas or details about patch size, augmentation and 2.5D ResUnet backbone [2,5] at the segmentation stage.
- The strong data augmentation is used in the model training process, such as random crop, random flip, etc., and a post-processing method of the pseudo-labels on hrT2 modality to improve the final output results.
- The experiments demonstrate the effectiveness of our proposed method.

2 Methods

The proposed cross-modal segmentation framework mainly consists of unsupervised domain adaption module for translating MR images from ceT1 to hrT2 modality, and semantic segmentation module for the VS and cochlea based on hrT2 scans and pseudo-labels of target domain. Domain Adaptation (DA) has recently raised strong interest in the medical imaging community. By encouraging algorithms to be robust to unseen situations or different data domains, domain adaptation has improved the applicability of deep learning approaches to various clinical scenes. In this challenge, we adopt CUT model [10] to realize the image-to-image translation. For the second stage, the semantic segmentation of organs or lesions is a common task in the field of medical image analysis.

There are already a large number of excellent and efficient algorithms available for medical image segmentation, such as U-Net [11], ResU-Net [2], nnU-Net [5], and others. Based on the strong baseline [4] and the strong heterogeneity of MRI images resolution, the optimal 2.5D ResUnet model is used as the segmentation workflow backbone to avoid the up/down-sample along z axis.

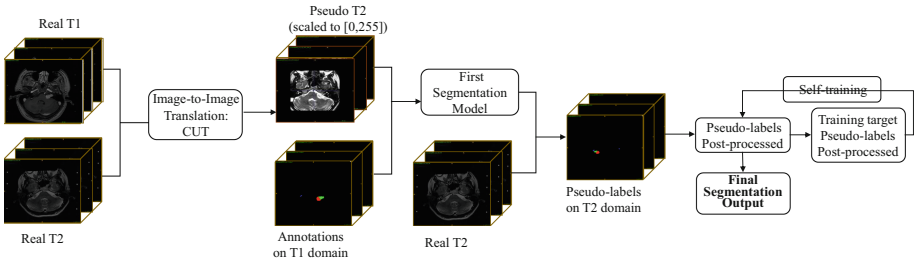

Fig. 1. An overview of our proposed cross-modal segmentation framework

2.1 Preprocessing

The preprocessing steps are listed as follows:

- Cropping strategy:
  Since the segmentation targets (the VS and cochlea) are located in the center of the brain MR image, we crop the center area as the regions of interest (ROIs), i.e., the inputs of our framework. According to the cropping method in reference [3], the 2D ROI images in the image-to-image translation stage are cropped with a range $[\frac{3W}{16} : \frac{13W}{16}, \frac{3H}{16} : \frac{13H}{16}]$ from the original range $[0 : W, 0 : H]$, and the 3D ROI images in the segmentation stage are cropped with a range $[0 : D, \frac{3W}{16} : \frac{13W}{16}, \frac{3H}{16} : \frac{13H}{16}]$ from the original range $[0 : D, 0 : W, 0 : H]$. Especially for the incomplete MRI images such as missing partial areas and irregular shape along x/y axis, the original images without any cropping are used as ROIs.
- Resampling method for anisotropic data:
  The 2D ROI images at the image-to-image translation stage are resampled to 256×256. The 3D ROI images at the segmentation stage are resampled to $D \times 384 \times 384$, the cube is randomly cropped along the z-axis, and the cube size is $40 \times 384 \times 384$.
- Intensity normalization method:
  All 2/3D ROI images are normalized to the range $[-1, 1]$ by the min-max scaler.
- Others:
  To improve the efficiency of the model training and inference, the mixed precision approach is used in the entire process of our framework.

2.2 Proposed Method

Our proposed framework is shown in Fig. 1, which mainly consists of the image-to-image translation and semantic segmentation of tumor and cochlea. The details of two stages are addressed as follows.

The Image-to-Image Translation. We apply the CUT model [10] to learn the mapping from the source domain ceT1 scans to the target domain hrT2 scans, which is similarly composed of the generator and discriminator. It is noteworthy that the discriminator has a stronger learning ability than the generator on this translation task, which implies the lower initial learning rate of the discriminator than generator could be more appropriate. After multiple repeated experiments, the CUT model could well implement the style transfer from the ceT1 scans to hrT2 scans with the following hyper-parameters setting: instance normalization, batch size equals to 10, the images range from -1 to 1, the initial learning late of generator is 0.0002, the initial learning rate of discriminator is 0.1×0.0002, and the proper stopping epoch. Additionally, we did not explore more details about the structure optimization of the CUT model due to the urgent time constraint. The visualization examples are shown in Fig. 2.

Semantic Segmentation. In the semantic segmentation stage, self-training is applied to further improve the segmentation results of the unseen real hrT2 scans. We adopt 2.5D ResUnet as the segmentation backbone, and network architecture has 3 down-sample layers, 3 up-sample layers, and no down-sample at z direction for the high performance and high efficiency of our method, which is shown in Fig. 3. The self-training process consists of the following steps.

- Step 1: The initial segmentation model is trained on the translated hrT2 scans and annotations of the ceT1 scans;
- Step 2: Generate pseudo-labels of real hrT2 scans by the segmentation model;
- Step 3: Select pseudo-labels by the max probability threshold p, and perform post-processing for the pseudo-labels to improve the segmentation quality of intra-meatal VS and cochlea;
- Step 4: Retrain the segmentation model on the selected real hrT2 scans and pseudo-labels;
- Step 5: Repeating Step 2–4, output the segmentation result, and evaluate it on the validation set.

Post-processing method for the pseudo-labels: we follow [12] to improve pseudo-labels of cochlea and intra-meatal VS based on the contrast of real hrT2 scans, and the examples are shown in Fig. 4.

Loss function: we use the summation between Dice loss and Cross-Entropy loss because compound loss functions have been proven to be robust in various medical image segmentation tasks [8].

Other tricks: 1) hard example mining is used in the loss function, and the hard samples are obtained by eroding or dilating the target masks; 2) to tackle

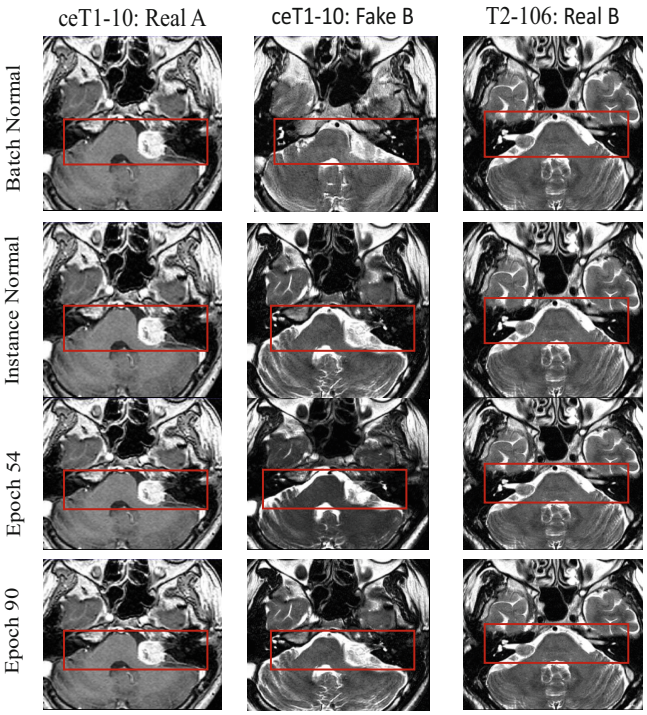

Fig. 2. Visualization for the transferred ceT1 scans and real ceT1/hrT2 scans of different hyper-parameters

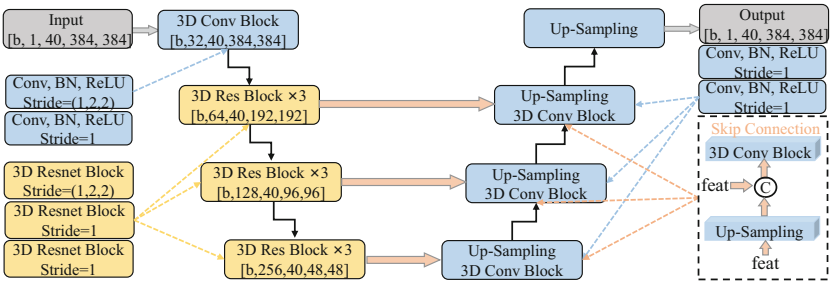

Fig. 3. The proposed network architecture

the structure heterogeneity of multi-center data set, such as the incomplete MRI cases with missing partital areas, the normal images are randomly replaced by the fixed ratio along x/y axis with the value -1, and the incomplete abnormal cases are randomly expanded to balance size x/y.

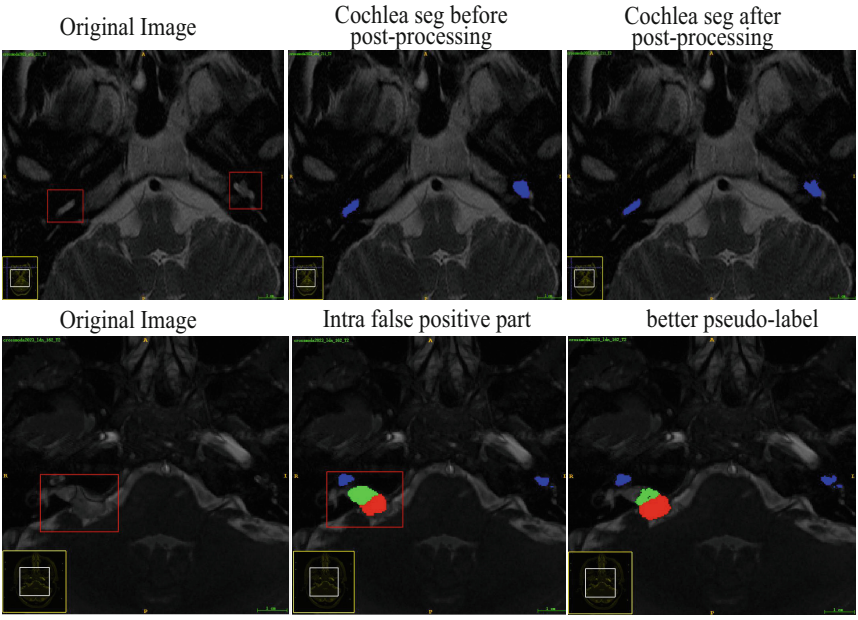

Fig. 4. Visualization for the pseudo-labels after post-processing

3 Results

3.1 Dataset and Evaluation Measures

The crossMoDA challenge 2023 organizer has publicly released 618 MRI scans, consisting of 227 ceT1 scans on the training source domain, 295 hrT2 scans on training target domain and 96 hrT2 scans for validation. The volumetric Dice coefficient, Average Symmetric Surface Distance (ASSD) and boundary ASSD [15] are used to evaluate algorithms.

3.2 Implementation Details

Environment Settings. The development environments and requirements are presented in Table 1.

Table 1. Development environments and requirements

Windows/Ubuntu version	Ubuntu 18.04.06 LTS
CPU	Intel(R) Xeon(R) Silver 4210R CPU @ 2.40 GHz
RAM	128 GB
GPU (number and type)	Two NVIDIA RTX A6000 48G
CUDA version	11.4
The programming language	Python 3.7
Deep learning framework	PyTorch (Torch 1.7.1+cu110, torchvision 0.8.2)
(Optional) Link to code	https://github.com/chencancan1018/crossMoDA

Data Augmentation. At the segmentation stage, random cropping ([0.8, 1.2]), flipping (z, y, z axis), elastic transforms (project MONAI [9]), contrast ([0,6, 1.5]), brightness ([0,6, 1.5]), and gamma augmentation ([0,6, 1.5]) are all applied to images, and contrast ([0,6, 1.5]), brightness ([0,6, 1.5]), and gamma augmentation ([0,6, 1.5]) are applied again to targets in the training process.

Training Protocols. Details of our training protocols are shown in Table 2 and Table 3.

Table 2. Training protocols for image-to-image translation

Network initialization	"he" normal initialization
Batch size	10
Patch size	256 × 256
Total epochs	200
Optimizer	ADAM ($weightdecay = 1e-4$)
Initial learning rate (lr)	2 × 1e−4, 0.2 × 1e−4
Lr decay schedule	plateau

Table 3. Training protocols for semantic segmentation

Network initialization	"he" normal initialization
Batch size	10
Patch size	40 × 384 × 384
Total epochs	200
Optimizer	ADAMW [7] ($weightdecay = 1e-4$)
Initial learning rate (lr)	1e−4
Lr decay schedule	CosineAnnealing

3.3 Results on the Validation and Test Set

The final scores on the validation and test set are listed in Table 4. Besides, the average inference time of the validation set is 27 s, and the max GPU occupancy is 5413M.

Table 4. Results of the proposed cross-modal segmentation framework on the validation set

Set	Intra Dice	extra Dice	VS Dice	Cochlea Dice	
Val	0.7057 ± 0.0959	0.8363 ± 0.108	0.8358 ± 0.1025	0.8355 ± 0.0441	
Test	0.699	0.808	0.828	0.844	
Set	Intra ASSD	extra ASSD	VS ASSD	boundary ASSD	Cochlea ASSD
Val	0.5928 ± 0.6929	0.5023 ± 0.2054	0.5648 ± 0.64	4.7632 ± 37.23	0.2259 ± 0.1449
Test	0.581	0.593	0.562	1.985	0.207

4 Conclusion

Based on 2.5D ResUnet, we propose an efficient cross-modal segmentation framework for the segmentation of VS and cochlea on cross-modality MRI images. The experimental results indicate that our framework is effective, but the lack of structure optimization and innovation of the image-to-image translation model expanded the gap between pseudo images and target domain (hrT2 modality). Moreover, the disadvantage yielded by the image-to-image translation could not be compensated at the segmentation stage, and we will focus on the designing and optimization of the translation model such as the exploration of 3D structure or the supervision of segmentation head.

References

1. Coelho, D.H., Tang, Y., Suddarth, B., Mamdani, M.: MRI surveillance of vestibular schwannomas without contrast enhancement: clinical and economic evaluation. Laryngoscope (2018)
2. Diakogiannis, F.I., Waldner, F., Caccetta, P., Wu, C.: ResUNet-a: a deep learning framework for semantic segmentation of remotely sensed data. ISPRS J. Photogramm. Remote. Sens. **162**, 94–114 (2020)
3. Dong, H., Yu, F., Zhao, J., et al.: Unsupervised domain adaptation in semantic segmentation based on pixel alignment and self-training. arXiv preprint arXiv:2109.14219 (2021)
4. Dorent, R., Kujawa, A., Ivory, M., Bakas, S., et al.: Crossmoda 2021 challenge: benchmark of cross-modality domain adaptation techniques for vestibular schwannoma and cochlea segmentation. Med. Image Anal. (2023)
5. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnU-net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021)

6. Kujawa, A., Dorent, R., Connor, S., Thomson, S., et al.: Deep learning for automatic segmentation of vestibular schwannoma: Dstudy from multi-centre routine mri. MedRxiv (2023)
7. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
8. Ma, J., Chen, J., Ng, M., Huang, R., Li, Y., et al.: Loss odyssey in medical image segmentation. Med. Image Anal. **71**, 102035 (2021)
9. Nic, M., Wenqi, L., Richard, B., Yiheng, W., Behrooz, H.: MONAI. https://github.com/Project-MONAI/MONAI. [Version 0.8.1]
10. Park, T., Efros, Alexei, A., Zhang, R., Zhu, J.Y.: Contrastive learning for unpaired image-to-image translation (2020)
11. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234–241 (2015)
12. Sallé, G., Conze, P.H., Bert, J., et al.: Tumor blending augmentation using one-shot generative learning for crossmodal MRI segmentation (2022). https://crossmoda-challenge.ml/media/papers-2022/latim.pdf
13. Shapey, J., Kujawa, A., Dorent, R., Wang, G., Dimitriadis, A., et al.: Segmentation of vestibular schwannoma from MRI an open annotated dataset and baseline algorithm. medRxiv (2021)
14. Shin, H., Kim, H., Kim, S., et al.: COSMOS: cross-modality unsupervised domain adaptation for 3D medical image segmentation based on target-aware domain translation and iterative self-training. arXiv preprint arXiv:2203.16557 (2022)
15. Wijethilake, N., Kujawa, A., Dorent, R., Asad, M., et al.: Boundary distance loss for intra-/extra-meatal segmentation of vestibular schwannoma. In: International Workshop on Machine Learning in Clinical Neuroimaging (2022)
16. Zhu, J.Y., Park, T., Isola, P., Efros, Alexei, A.: Unpaired image-to-image translation using cycle-consistent adversarial networks, pp. 2223–2232 (2017)