MS-MT++: Enhanced Multi-Scale Mean Teacher for Cross-Modality Vestibular Schwannoma and Cochlea Segmentation

Ziyuan Zhao[1,2], Ruikai Lin[1,3], Kaixin Xu[1], Xulei Yang[1], and Cuntai Guan[2]

[1] Institute for Infocomm Research (I^{2}R), A*STAR, Singapore
[2] Nanyang Technological University, Singapore
[3] National University of Singapore, Singapore

Abstract. Domain shift has been a long-standing issue for medical image segmentation. Unsupervised domain adaptation (UDA) methods have recently achieved promising cross-modality segmentation performance by distilling knowledge from a label-rich source domain to a target domain without labels. Different from CrossMoDA 2022, the challenge this year includes highly heterogeneous MRI scans from more institutions and various scanners and subdivides the segmentation object into three key brain structures (intra/extra-vestibular schwannoma and cochlea). In this work, we improve our previous method and propose an enhanced multi-scale self-ensembling-based UDA framework for automatic segmentation of Vestibular Schwannoma and Cochlea on high-resolution T2 images. Our method demonstrated a mean Dice score of 0.5679 and 0.7293 for the extra/intra-VS joint area and Cochlea respectively, secured a top 10 finish in the validation phase of the CrossMoDA 2023 challenge.

Keywords: Medical image segmentation · Unsupervised domain adaptation · Vestibular Schwannoma · Cochlea

1 Introduction

Medical image segmentation plays a pivotal role in medical image analysis, providing crucial information for diagnostic analysis and treatment planning [5]. The accurate segmentation and measurement of Vestibular Schwannoma (VS) and Cochlea from MRI are instrumental in assisting VS treatment planning and streamlining clinical workflows [13]. As a result, researchers have increasingly turned to deep learning as a solution for autonomously segmenting VS and Cochlea. While Contrast-enhanced T1 (ceT1) MR imaging is commonly used, there is a growing interest in non-contrast sequences, such as high-resolution T2 (hrT2) imaging. Recent findings indicate that high-resolution T2 (hrT2) MRI could be a safer and more cost-efficient alternative to contrast-enhanced T1 (ceT1) MRI. However, the significant domain shift between MRI images with different contrasts, along with the expensive and laborious process of re-annotating medical image scans in another modality, makes it challenging for deep learning to generalize effectively across both domains. In response, unsupervised domain

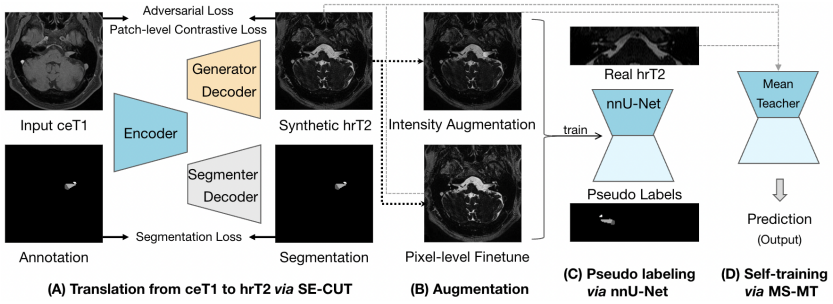

Fig. 1. The workflow of our proposed method. First, the SE-CUT network is proposed to translate ceT1 to hrT2. Then, the target areas are augmented with intensity, and multiple CycleGANs are used for pixel-level enhancement. After that, all synthetic and augmented scans are used for training a 3D nnU-Net, which can generate pseudo labels for all unlabeled real hrT2 images. Finally, an MS-MT network is employed for self-ensemble learning.

adaptation (UDA) has emerged as a promising approach in medical imaging to enhance the robustness of machine learning algorithms, enabling them to adapt to diverse input data domains and extending their utility in various clinical scenarios. Therefore, we are encouraged to perform unsupervised domain adaptation (UDA) and conduct VS and Cochlea segmentation in the hrT2 domain by leveraging both labeled ceT1 scans and unlabeled hrT2 scans.

In this work, we proposed an effective cross-modality UDA framework. In our framework, we first translate ceT1 scans to hrT2 modality via a segmentation-enhanced unpaired contrastive unpaired translation (SE-CUT) network. Then, three CycleGAN networks are trained for pixel-level intensity fine-tuning. We also perform intensity augmentation on the annotated regions of generated images. Then we apply a 3D full resolution nnU-Net to generate pseudo labels for unlabeled real hrT2 scans. Finally, we build a multi-scale mean-teacher (MS-MT) network [15, 8] for improving the cross-modality segmentation performance. The experimental results show that the proposed UDA network can greatly reduce the domain gap, achieving promising segmentation performance on ceT2 scans.

2 Methods

Given an unpaired dataset of two modalities, *i.e.,* annotated ceT1 MRI images $\mathcal{D}_s = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^N$ and non-annotated hrT2 MRI scans $\mathcal{D}_t = \{(\mathbf{x}_i^t)\}_{i=1}^M$, sharing the same classes (intra- and extra-meatal VS, and Cochlea), we aim to exploit $\mathcal{D}_s$ and $\mathcal{D}_t$ for unsupervised domain adaptation to enhance the cross-modality segmentation performance of the VS and Cochlea on hrT2 MRI images. The overview of our UDA framework is shown in Fig. 1.

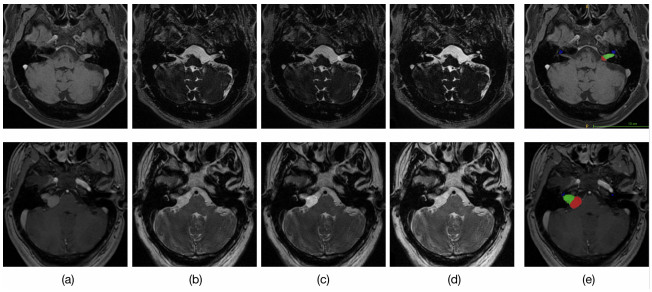

Fig. 2. Visual comparison for (a) Original ceT1, (b) Generated hrT2 using SE-CUT, (c) Generated hrT2 after intensity augmentation, (d) Generated hrT2 after pixel-level fine-tuning via CycleGAN, and (e) Ground truth of VS and cochlea.

2.1 Segmentation-enhanced translation

To address the domain gap between modalities, we conduct image-level domain adaptation to create synthetic target samples. This involves training a model on synthetic target images for VS and Cochlea segmentation in real hrT2 scans. For image-to-image translation, we use the Contrastive Unpaired Translation (CUT) [12] method for time efficiency. To preserve structural information during translation (see Fig. 1), we enhance the 2D CUT with an additional segmentation decoder. This ResNet-based generator converts source domain images to the target domain, while a PatchGAN discriminator distinguishes real from generated images [12]. Using the SIFA architecture [1], we connect two layers of the encoder with the segmenter decoder to produce multi-level segmentation predictions. This segmentation loss guides the encoder to focus on relevant areas, preserving the structure of VS and Cochlea in the translated images.

2.2 Intensity augmentation and pseudo labeling

Considering that tumors exhibit T2 heterogeneous signal intensity [11] and cochleas show T2 hyperintense signal intensity [9], we perform intensity augmentation and generate augmented data for diversifying the training distributions. Using the generated hrT2 images and the ground truth annotations, the signal intensities of annotation regions for each scan are randomly selected by a factor of 50% or 25% for mute and intensify respectively, doubling the training data. Meanwhile, we train multiple CycleGANs [19] for different sources of data for further fine-tuning in pixel-level of synthetic hrT2 to improve the accuracy of the data and the effectiveness of the training, which also augmented the data by another copy. On the other hand, to enhance segmentation performance on real hrT2 images, we utilize a Pseudo-Labeling (PL) strategy to leverage unlabeled hrT2 images by generating pseudo hrT2 annotations. We train a 3D full-resolution nnU-Net [6] using synthetic hrT2 images and augmented images. These trained models are then used to generate pseudo-labels for the unlabeled hrT2 images, further improving segmentation accuracy.

2.3 Multi-scale self-ensembling learning

To fully utilize available data, we propose using the self-ensembling network, mean teacher (MT) [15], where a teacher model is constructed with the same architecture as the student model and updated using an exponential moving average (EMA) of the student's parameters during training. We ensure consistency between student and teacher outputs by minimizing their difference with the mean square error (MSE) loss. Additionally, we leverage the success of multi-scale learning in medical image analysis [4, 18, 8, 17] and introduce a multi-scale mean teacher (MS-MT) network following [8]. This approach utilizes multi-scale predictions for deep supervision and consistency regularization. Both the teacher and student networks use the 3D full-resolution nnU-Net [6] as our backbone, with auxiliary layers connected to each block of the last five blocks to obtain multi-scale predictions. This combination allows us to make the most of available data for improved segmentation performance.

3 Experiments and Results

The CrossMoDA 2023 challenge provides a highly heterogeneous dataset from multiple centers, comprising 227 annotated ceT1 scans and 391 unlabeled hrT2 scans (with 295 scans used as training targets and 96 scans used as the validation set) [3, 7, 16]. The London (LDN) and Tilburg SC-GK (ETZ) data were obtained using different scanners and imaging sequences, with both T1-weighted and high-resolution T2-weighted imaging performed. The UK MC-RC (UKM) data were obtained on various scanners with different magnetic field strengths, and slice thickness, voxel volume and intensities vary significantly across all ceT1 weighted and T2 weighted imaging. As the voxel spacing varied in the raw scans, the images were resampled into a common spacing of $0.6 \times 0.6 \times 1.0$ mm and the intensity was normalized to [0, 1] using Min-Max scaling. To get the regions of interest (ROI) and remove the noise, the images were cropped into 256×256 pixels in the xy-plane using a 75-percentile binary threshold [2], resulting in $256 \times 256\times$ N image volumes for 3D nnU-Net training. These processed 3D vol-umes were also sliced along the z-axis to create N number of 2D images for SE-CUT and CycleGAN training. The Dice Score (DSC [%]) [14] and the Average Symmetric Surface Distance(ASSD [voxel]) [10] are used to assess the model performance on VS and Cochlea segmentation.

We used a single NVIDIA A40 GPU with 48GB of memory for model training. Fig. 2 shows the effects of using SE-CUT to translate (a) original ceT1 into (b) synthetic hrT2 after training for 120 epochs. In the SE-CUT, We followed [12] and [1] to keep the same weights for adversarial loss, contrastive loss, and segmentation loss; the loss weights for additional segmentation were set to 1 and 0.1 for the last layer and the second last downsampling layer, respectively. Structural information for VS and cochlea in the original ceT1 scans was better retained because of the segmentation-enhanced architecture. Figures 2 (c) and (d) demonstrate the results of synthetic hrT2 images after intensity augmentation and pixel-level fine-tuning, respectively. Three CycleGANs were trained

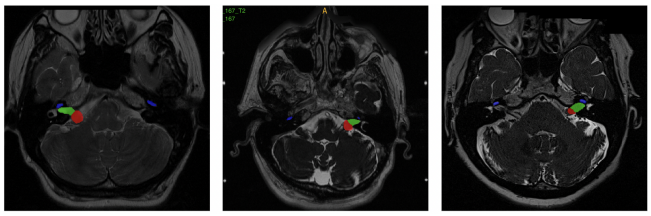

Fig. 3. Segmentation results of the validation set produced by our workflow. The intra-VS, extra-VS, and cochlea are indicated in red, green, and blue color, respectively.

with 50 epochs and performed on scans from different sources (ETZ, LDN, and UKM, respectively) to bring the pixel intensity closer to the real hrT2 scans from the corresponding source. The three sets of data were collectively used to train a 3D full resolution nnU-Net with generalization ability for the hrT2 modality. After 400 epochs of training and 5-fold cross validation, the nnU-Net was used to generate pseudo labels for all unlabeled real hrT2 scans. Then we performed the self-training process using synthetic hrT2 with real labels and real hrT2 with pseudo labels. The MS-MT network with the backbone of two 3D nnU-Nets was trained with an initial learning rate of 0.01 using stochastic gradient descent for 300 epochs. The combination of Dice and cross-entropy losses was used as the objective function, and the deep supervision scheme in nnU-Net [6] was also enabled. In the MS-MT network, the EMA update α was set to 0.9, and the loss weights for consistency regularization were set to $\{0.05, 0.05, 0.05, 0.4, 0.5\}$, assigned to each feature map according to its size in ascending order. After performing this step, the validation set was inferred by the 5-fold ensemble MS-MT model (without post-processing), as shown in Figure 3, and the Dice score of VS and cochlea in the validation phase leaderboard were 0.7293 and 0.5679, respectively. Finally, the largest connected component (LCC) was calculated, and the segmentation results were post-processed (*i.e.*, the first LCC was preserved for the connected intra and extra VS while the first and second LCCs were preserved for the cochlea) for final submission. Fig. 3 highlights some qualitative results produced by our method. Due to the page limit, ablation analysis and other exploration studies are not included in the short paper.

4 Conclusion

In this work, we propose a four-stage cross-modality unsupervised domain adaptation workflow, including unpaired image translation, augmentation and pixel-level fine-tuning, pseudo labeling, and self-training. A final model for segmenting intra-meatal VS, extra-meatal VS and cochlea on hrT2 scans was achieved with only annotated ceT1 scans, bridging the gap between the two different domains. Finally, we achieved a top 10 finish in the validation phase, demonstrating its effectiveness in bridging the gap between the ceT1 and hrT2 MRI modalities in the CrossMoDA 2023 challenge.

References

1. Chen, C., Dou, Q., Chen, H., Qin, J., Heng, P.A.: Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation. IEEE transactions on medical imaging **39**(7), 2494–2505 (2020)
2. Choi, J.W.: Using out-of-the-box frameworks for unpaired image translation and image segmentation for the crossmoda challenge. arXiv e-prints pp. arXiv–2110 (2021)
3. Dorent, R., Kujawa, A., Ivory, M., Bakas, S., Rieke, N., Joutard, S., Glocker, B., Cardoso, J., Modat, M., Batmanghelich, K., et al.: Crossmoda 2021 challenge: Benchmark of cross-modality domain adaptation techniques for vestibular schwannoma and cochlea segmentation. arXiv preprint arXiv:2201.02831 (2022)
4. Dou, Q., Yu, L., Chen, H., Jin, Y., Yang, X., Qin, J., Heng, P.A.: 3d deeply supervised network for automated segmentation of volumetric medical images. Medical image analysis **41**, 40–54 (2017)
5. Hesamian, M.H., Jia, W., He, X., Kennedy, P.: Deep learning techniques for medical image segmentation: achievements and challenges. Journal of digital imaging **32**(4), 582–596 (2019)
6. 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. Nature methods **18**(2), 203–211 (2021)
7. Kujawa, A., Dorent, R., Connor, S., Thomson, S., Ivory, M., Vahedi, A., Guilhem, E., Bradford, R., Kitchen, N.D., Bisdas, S., Ourselin, S., Vercauteren, T.K.M., Shapey, J.: Deep learning for automatic segmentation of vestibular schwannoma: A retrospective study from multi-centre routine mri. In: medRxiv (2022)
8. Li, S., Zhao, Z., Xu, K., Zeng, Z., Guan, C.: Hierarchical consistency regularized mean teacher for semi-supervised 3d left atrium segmentation. In: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). pp. 3395–3398. IEEE (2021)
9. Lin, E., Crane, B.: The management and imaging of vestibular schwannomas. American Journal of Neuroradiology **38**(11), 2034–2043 (2017)
10. Lu, F., Wu, F., Hu, P., Peng, Z., Kong, D.: Automatic 3d liver location and segmentation via convolutional neural network and graph cut. International journal of computer assisted radiology and surgery **12**(2), 171–182 (2017)
11. Nguyen, D., de Kanztow, L.: Vestibular schwannomas: a review. Appl Radiol **48**(3), 22–27 (2019)
12. Park, T., Efros, A.A., Zhang, R., Zhu, J.Y.: Contrastive learning for unpaired image-to-image translation. In: European conference on computer vision. pp. 319–345. Springer (2020)
13. Shapey, J., Wang, G., Dorent, R., Dimitriadis, A., Li, W., Paddick, I., Kitchen, N., Bisdas, S., Saeed, S.R., Ourselin, S., 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), 171–179 (2019)
14. Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Cardoso, M.J.: Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Deep learning in medical image analysis and multimodal learning for clinical decision support, pp. 240–248. Springer (2017)

15. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. arXiv preprint arXiv:1703.01780 (2017)
16. Wijethilake, N., Kujawa, A., Dorent, R., Asad, M.H., Oviedova, A., Vercauteren, T., Shapey, J.: Boundary distance loss for intra-/extra-meatal segmentation of vestibular schwannoma. In: MLCN@MICCAI (2022)
17. Zhao, Z., Hu, J., Zeng, Z., Yang, X., Qian, P., Veeravalli, B., Guan, C.: Mmgl: Multi-scale multi-view global-local contrastive learning for semi-supervised cardiac image segmentation. arXiv preprint arXiv:2207.01883 (2022)
18. Zhao, Z., Zeng, Z., Xu, K., Chen, C., Guan, C.: Dsal: Deeply supervised active learning from strong and weak labelers for biomedical image segmentation. IEEE Journal of Biomedical and Health Informatics **25**(10), 3744–3751 (2021)
19. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)