Unsupervised Domain Adaptation in Semantic Segmentation Based on Pixel Alignment and Self-Training (PAST)

Hexin Dong[1], Fei Yu[1], Mingze Yuan[1], Jie Zhao[3], Bin Dong[4,1], and Li Zhang[1,2⋆]

[1] Center for Data Science, Peking University, Beijing, China
[2] Center for Data Science in Health and Medicine, Peking University, Beijing, China
[3] National Engineering Laboratory for Big Data Analysis and Applications, Peking University, Beijing, China
[4] Beijing International Center for Mathematical Research (BICMR), Peking University, Beijing, China
{donghexin,yufei1900,mzyuan,jiezhao,zhangli_pku}@pku.edu.cn
dongbin@math.pku.edu.cn

Abstract. This paper proposes an unsupervised cross-modality domain adaptation approach based on pixel alignment and self-training. Pixel alignment transfers the scans in ceT1 to that in hrT2, helping to reduce domain shift in the training segmentation model. Self-training adapts the decision boundary of the segmentation network to fit the distribution of hrT2 scans. Experiment results show that PAST has outperformed the non-UDA baseline significantly, and it received rank two on the CrossMoDA2022 validation phase Leaderboard with a mean Dice score of 0.8511.

Keywords: unsupervised domain adaptation · pixel alignment · self-training.

1 Introduction

CrossModa challenge [6,7,8] aims to segment two types of critical intracranial objects involved in the treatment planning of vestibular schwannoma (VS): the tumor and cochlea. While contrast-enhanced T1 (ceT1) Magnetic Resonance Imaging (MRI) scans are commonly used for VS segmentation, recent work has demonstrated that high-resolution T2 (hrT2) imaging could be a reliable, safe, and lower-cost alternative to ceT1. Therefore, the challenge participants are asked to provide a segmentation model of VS and cochlea on hrT2 scans based on unsupervised domain adaptation (UDA) using only the information of labeled ceT1 scans and unlabeled hrT2 scans.

To solve this problem, we propose an effective and intuitive UDA method combining pixel-level alignment and self-training (PAST). Firstly, we transfer labeled images from the ceT1 domain to the hrT2 domain so that images can

⋆ Correspondence to: zhangli_pku@pku.edu.cn

be aligned into the same distribution. Secondly, the model is further trained on pseudo labels generated from transferred ceT1 scans and hrT2 scans, which find a better decision boundary on the hrT2 domain. The experimental results show that our method greatly reduces the domain shift and achieves 2nd place with a dice score of 0.8511 on the validation set.

2 Methods and Experimental Methods

2.1 Method Overview

We introduce our method in this section. Our method has two major parts: pixel-level alignment and the self-training stage.

First, we follow [2] to learn a mapping from the source domain to the target domain, i.e., we transfer ceT1 scans to hrT2 scans. After doing so, we can use synthesized hrT2 scans to train a segmentation model using supervised learning. As shown in Figure 2, the model achieves the domain adaptation using NiceGAN [1] (i.e., an extension method of CycleGAN), which reuses discriminators for encoding to improve the efficiency and effectiveness of training.

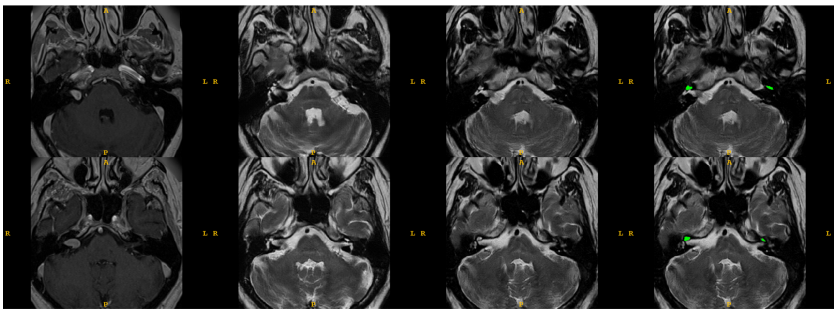

Fig. 1. Visualization for transferred ceT1 scans. From left to right: (1) ceT1 scans. (2) synthesized hrT2 scans without segmentor. (3) synthesized hrT2 scans with segmentor. (4) cochlea ground truth.

Different from the model we have used in CrossmModa2021[11], we follow [12] and add an extra segmentor to segment transferred ceT1 scans. As shown in Figure 1, the segmentor helps preserve the detailed structures, especially for small and inconspicuous cochlea.

Second, we apply self-training to further improve the decision boundary of the segmentation model. Similar to [5], we introduce a super parameter q of the pixel portion. We iteratively generate the pseudo label $\hat{y_c}$ using the top q of pixels in segmentation output y_c with a higher probability of retraining the model. The overall training process of the proposed method is summarized in Algorithm 1.

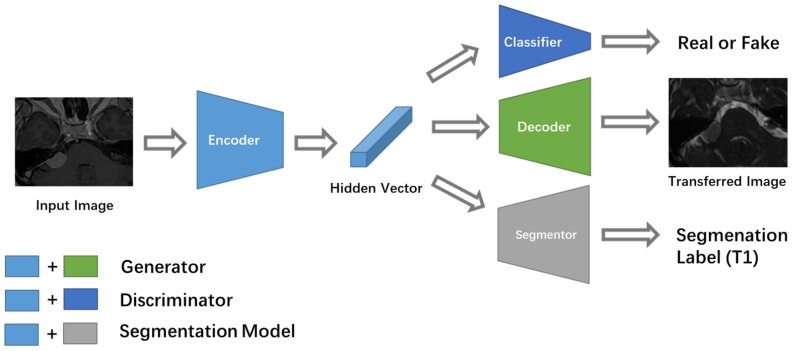

Fig. 2. Flowchart of NiceGAN [1]. It extracts features from the input image with the shared Encoder. The Classifier from the Discriminator distinguishes the real or fake feature vectors. The Decoder from the Generator generates transferred images. The Segmentor segments transferred ceT1 scans with a 2d label.

All models are implemented using PyTorch 1.9. Pixel-level alignment model runs on a single V100 GPU with 16 GB memory, and the self-training model runs on a single TIAN V GPU with 12 GB memory. All training data are collected from CrossModa training set [8,6], and we verify our model on the CrossModa2022 validation set.

Algorithm 1 training process of the proposed method

1: Initialize ceT1 scans images and label (X_s, y_s), hrT2 scans images X_t, Segmentation network S, Image translation network T
2: Train network T with X_s and X_t
3: Transfer ceT1 scans X_s to $\hat{X}_s$ using T
4: Train network S with $(\hat{X}_s, y_s)$
5: Initialize concat scans images $X_c = \{\hat{X}_s, X_t\}$, self-training segmentation network $S_0 = S$
6: **for** $k \leftarrow 1$ to K **do**
7: input X_c into S_{k-1} and generate pseudo label $\hat{y_c^k}$ with a fixed portion q_k
8: Initialize $S_k \leftarrow S_{k-1}$
9: Train S_k with $(X_c, \hat{y_c^k})$
10: **end for**
11: **return** S_k

2.2 Experiments

For the preprocessing step, we observe that the segmentation targets are located in the center of the image, so we take the center area of the image as the region of interest (ROI). For the image translation, the 2D images with a range of

$[0:W, 0:H]$ are cropped into the 2D ROI with a range of $[\frac{3W}{16}:\frac{13W}{16}, \frac{3H}{16}:\frac{13H}{16}]$. For the 3D segmentation, the volumetric data ($[0:W, 0:H, 0:D]$) will be cropped into the 3D ROI with a range of $[\frac{3W}{16}:\frac{13W}{16}, \frac{3H}{16}:\frac{13H}{16}, 0:D]$. After then, the intensity values in ROI are normalized by rescaling to $[0-255]$

In the pixel alignment stage, we adopt NICE-GAN [1] on 2D transverse slides of the ROIs, transferring ceT1 to hrT2. We then concatenate the synthesized 2D hrT2 slides to a 3D volumetric image. For 3D segmentation, we follow the nnUNet framework [3]. Several research settings are implemented. We select nnUnet and ResUNet(i.e., an extended version for nnUnet with ResNet encoder) as our segmentation model. First, we train the models with paired synthesized ceT1 scans and labels. We named the above two models **nnUNetPA** and **ResUNetPA**. The self-training stage generally follows the Algorithm 1. The nnUnets (nnUNetPA and ResUNetPA) are used as pretrain models for self-training, i.e. S_0 in Algorithm 1. The synthesized images and pseudo labels are derived from the corresponding image-to-image translation models and nnUnets, respectively. In practice, we set the initial q to 0.6, the maximum iteration K of self-training to 2, and the initial learning rate to 0.008. We name these model as **nnUNetPAST2** and **ResUNetPAST2**.

As shown in Table 1, our models perform well on cochlea while having some problems on VS. We notice a appearance gap on VS between London data (CrossModa 2021 cases) and Tilburg data(CrossModa 2022 cases). The gap causes mode collapse on our generation model and performance drop on segmenting the VS. Our model achieved the previous SOTA on VS on CrossModa2021[13] (we name this model as **PAST1.0**). However, the results shown in Table 2 indicate that PAST1.0 performs not so well on Tilburg data due to a small domain gap between these two datasets. To solve this problem, we incrementally train PAST1.0 with two extra self-training stages (with only hrT2 training data) and name it as **IResUNetPAST2**. As shown in Table 2, IResUNetPAST2 improves the performance on Tilburg dataset. Finally, we combine the above models (i.e., PAST1.0 to segment VS on the London dataset, IResUNetPAST2 to segment VS on the Tilburg dataset, and nnUNetPAST2 to segment cochlea) and achieve the best results. We thus name it as **PAST2.0**, raising the overall Dice to 0.8511.

Table 1. Segmentation results for selected model.

Model Name	VS Dice	Cochlea Dice	Mean Dice
nnUnetPA	0.6716 ± 0.2564	0.8280 ± 0.0306	0.7498 ± 0.1306
ResUNetPA	0.6729 ± 0.2533	0.8246 ± 0.0294	0.7487 ± 0.1284
nnUNetPAST2	0.8095 ± 0.0960	0.8547 ± 0.0283	0.8320 ± 0.0716
ResUNetPAST2	0.8115 ± 0.0977	0.8515 ± 0.0297	0.8315 ± 0.0848
PAST1.0	0.7935 ± 0.2029	0.7677 ± 0.0525	0.7806 ± 0.1133
IResUNetPAST2	0.8381 ± 0.0794	0.8412 ± 0.0249	0.8386 ± 0.0774
PAST2.0(47.0)	0.8473 ± 0.0633	0.8547 ± 0.0283	0.8511 ± 0.0322

Table 2. VS Segmentation results for selected model.

Model Name	London data Dice	Tilburg data Dice	Mean Dice
nnUNetPAST2	0.8231 ± 0.1129	0.7959 ± 0.0728	0.8095 ± 0.0960
ResUNetPAST2	0.8281 ± 0.1049	0.7949 ± 0.0742	0.8115 ± 0.0848
PAST1.0	0.8705 ± 0.0646	0.7170 ± 0.2554	0.7935 ± 0.2014
IResUNetPAST2	0.8519 ± 0.0976	0.8243 ± 0.0520	0.8381 ± 0.0794
PAST2.0	0.8705 ± 0.0646	0.8243 ± 0.0520	0.8474 ± 0.0629

References

1. Runfa Chen, Wenbing Huang, Binghui Huang, Fuchun Sun, and Bin Fang. Reusing discriminators for encoding: Towards unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8168–8177, 2020.
2. Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. PMLR, 2018.
3. Fabian Isensee, Paul F Jäger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. Automated design of deep learning methods for biomedical image segmentation. *arXiv preprint arXiv:1904.08128*, 2019.
4. Guoliang Kang, Yunchao Wei, Yi Yang, Yueting Zhuang, and Alexander G Hauptmann. Pixel-level cycle association: A new perspective for domain adaptive semantic segmentation. *arXiv preprint arXiv:2011.00147*, 2020.
5. Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6936–6945, 2019.
6. Jonathan Shapey, Aaron Kujawa, Reuben Dorent, Guotai Wang, Alexis Dimitriadis, Diana Grishchuk, Ian Paddick, Neil Kitchen, Robert Bradford, Shakeel R Saeed, Sotirios Bisdas, Sébastien Ourselin, and Tom Vercauteren. Segmentation of vestibular schwannoma from mri — an open annotated dataset and baseline algorithm. *medRxiv*, 2021.
7. Jonathan Shapey, Aaron Kujawa, Reuben Dorent, Guotai Wang, Alexis Dimitriadis, Diana Grishchuk, Ian Paddick, Neil Kitchen, Robert Bradford, Shakeel R Saeed, Sotirios Bisdas, Sébastien Ourselin, and Tom Vercauteren. Segmentation of vestibular schwannoma from mri — an open annotated dataset and baseline algorithm. *Scientific Data*, 2021. In press. Preprint available at medRXiv:10.1101/2021.08.04.21261588.
8. Kujawa A. Dorent R. Wang G. Bisdas S. Dimitriadis A. Grishchuck D. Paddick I. Kitchen N. Bradford R. Saeed S. Ourselin S. & Vercauteren T Shapey, J. Segmentation of vestibular schwannoma from magnetic resonance imaging: An open annotated dataset and baseline algorithm [data set]. *The Cancer Imaging Archive*, 2021.
9. Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12414–12424, 2021.

10. Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.
11. Hexin Dong, Fei Yu, Jie Zhao, Bin Dong and Li Zhang.:Unsupervised Domain Adaptation in Semantic Segmentation Based on Pixel Alignment and Self-Training. arXiv preprint arXiv:2109.14219
12. Hyungseob Shin, Hyeongyu Kim, Sewon Kim, Yohan Jun, Taejoon Eo, Dosik Hwang: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
13. Dorent, R. et al (2022). CrossMoDA 2021 challenge: Benchmark of Cross-Modality Domain Adaptation techniques for Vestibular Schwannoma and Cochlea Segmentation. ArXiv https://arxiv.org/abs/2201.02831