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

Hexin Dong[1*], Fei Yu[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 *These authors contributed equally
{donghexin,yufei1900,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 ceT1 scans to hrT2 modality, 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-2 on CrossMoDA validation phase Leaderboard with a mean Dice score of 0.8395.

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

1 Introduction

CrossModa challenge[5,6,7] aims to segment two critical brain structures involved in the treatment planning of vestibular schwannoma (VS): the tumor and the 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 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 be aligned into the same distribution. Secondly, the model is further trained on

⋆ Correspondence to: zhangli_pku@pku.edu.cn

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 0.8395 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 1, 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.

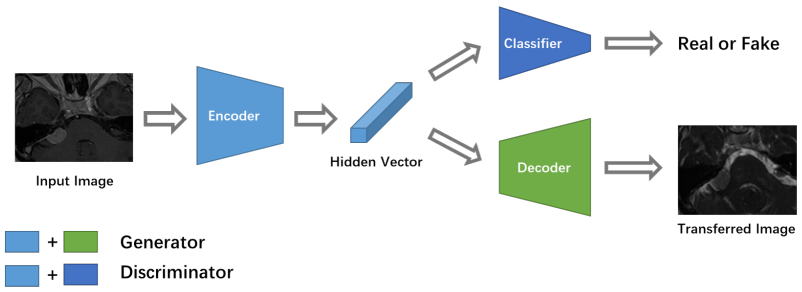

Fig. 1. flowchart of NiceGAN[1]. It extracts feature from the input image with the shared Encoder. The Classifer from the Discriminator distingguishes real or fake feature vectors. The Decoder from the Generator generates transferred images.

Second, we apply self-training to further improve the decision boundary of the segmentation model. Similar to [4], 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 to retrain the model. Overall training process of the proposed method is summarized in Algorithm 1.

All models are implemented using the PyTorch 1.9. Pixel-level alignment model runs on a single V100 GPU with 16 GB memory and self-training model runs on a single TIAN V GPU with 12 GB memory. All training data used are collected from CrossModa training set [7,5] and we verify our model on crossModa validation set.

Algorithm 1 training process of the proposed method

Initialize ceT1 scans images and label $(X_s; y_s)$, hrT2 scans images X_t, Segmentation network S, Image translation network T
Train network T with X_s and X_t
Transfer ceT1 scans X_s to $\hat{X}_s$ using T
Train network S with $(\hat{X}_s; y_s)$
Initialize concat scans images $X_c = \{\hat{X}_s; X_t\}$, self-training segmentation network $S_0 = S$
for $k \leftarrow 1$ to K **do**
    input X_c into S_{k-1} and generate pseudo label $\hat{y_c^k}$ with a fixed portion q_k
    Initialize $S_k \leftarrow S_{k-1}$
    Train S_k with $(X_c; \hat{y_c^k})$
end for
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) (Figure 2). 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 $[\frac{W}{4} : \frac{3W}{4}; \frac{H}{4} : \frac{3H}{4}]$. For the 3D segmentation, the volumetric data ($[0 : W; 0 : H; 0 : D]$) will be cropped into the 3D ROI with a range of $[\frac{W}{4} : \frac{3W}{4}; \frac{3H}{8} : \frac{3H}{4}; 0 : D]$. 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. First, we train the models with paired synthesized hrT2 scans and labels. Since most data have a dimension of 448 pixels, we thus call this model nnUnet448. However, we notice that there're two types of protocols in hrT2 with significantly different appearances (one with a dimension of 448 pixels, called 448 scans, and another with a dimension of 384 pixels, called 384 scans). We thus create a nnUnet384 model for those 384 scans. We evaluate the results of the two models on all data and the results of their respective applicable data. We call the model in the latter scenario as nnUnetCon.

The self-training stage generally follows the Algorithm 1. The nnUnets (nnUnet448, nnUnet384 and nnUnetCon) 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 is set to 0.08. We name this model as nnUetST2.

Apart from this, we train a ResUnet (i.e., an extended version for nnUnet with ResNet encoder) following the above stages and name this model as ResUnetST2. A combined version using nnUetST2 to segment cochlea and Re-

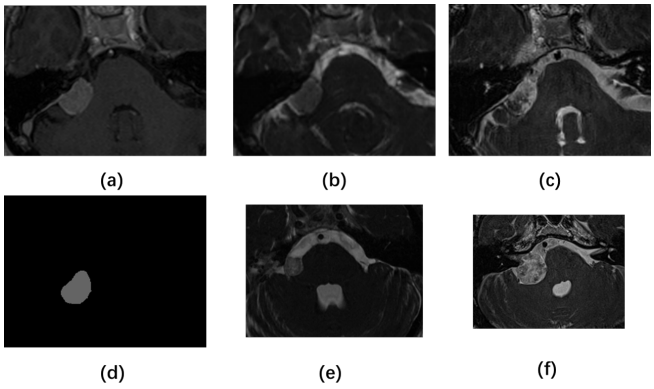

Fig. 2. Croped images samples: a).ceT1 sample b).transfered 448 ceT1 sample c).transfered 384 ceT1 sample d).label e).448 hrT2 sample f).384 hrT2 sample

sUnetST2 to segment VS has also been evaluated and achieves a better result. We thus name it the proposed PAST.

3 Results

Table 1 shows the results for the aforementioned models. Compared to nnUnet without DA, i.e., training with ceT1 scans directly, both nnUnet448 and nnUnet384 have a noticeable improvement, which shows the effectiveness of the pixel alignment. However, the two models did not achieve satisfactory accuracy because hrT2 modality itself has two different protocols. nnUnetCon model solves this problem and further improves the performance with model ensembling. Furthermore, the experiments show that self-training achieves better performance on overall Dice, but different network backbones behave differently on either VS or cochlea segmentation (Figure 3). Thus, as our final proposed method, PAST combines all the merits from the aforementioned network architectures and training techniques, raising the overall Dice to 0.8395.

Table 1. Segmentation results for selected model.

Model Name	VS Dice	Cochlea Dice	Mean Dice
nnUnet without DA	0:0549 ± 0:1859	0:1905 ± 0:1643	0:1227 ± 0:1192
nnUnet448	0:7509 ± 0:2683	0:7818 ± 0:0425	0:7664 ± 0:1417
nnUnet384	0:5905 ± 0:3754	0:7870 ± 0:0413	0:6887 ± 0:1929
nnUnetCon	0:8281 ± 0:1679	0:7949 ± 0:0332	0:8115 ± 0:0848
nnUnetST2	0:8553 ± 0:0871	0:8089 ± 0:0334	0:8321 ± 0:0435
ResUnetST2	0:8700 ± 0:0657	0:7820 ± 0:0295	0:8260 ± 0:0349
PAST	0:8700 ± 0:0657	0:8089 ± 0:0335	0:8395 ± 0:0328