Unsupervised Cross-Modality Domain Adaptation for Segmenting Vestibular Schwannoma and Cochlea with Data Augmentation and Model Ensemble

Hao Li[1], Dewei Hu[1], Qibang Zhu[1], Kathleen E. Larson[1], Huahong Zhang[1], and Ipek Oguz[1]

Vanderbilt University, Nashville, TN 37235, USA

Abstract. Magnetic resonance images (MRIs) are widely used to quantify vestibular schwannoma (VS) and the cochlea. Recently, deep learning methods have shown state-of-the-art performance for segmenting these structures. However, training segmentation models may require manual labels in target domain, which is expensive and time-consuming. To overcome this problem, domain adaptation is an effective way to leverage information from source domain to obtain accurate segmentations without requiring manual labels in target domain. In this paper, we propose an unsupervised learning framework to segment the VS and cochlea. Our framework leverages information from contrast-enhanced T1-weighted (ceT1-w) MRIs and its labels, and produces segmentations for T2-weighted MRIs without any labels in the target domain. We first applied a generator to achieve image-to-image translation. Next, we ensembled outputs from an ensemble of different models to obtain final segmentations. To cope with MRIs from different sites/scanners, we applied various 'online' data augmentations during training to better capture the geometric variability and the variability in image appearance and quality. Our method is easy to build and produces promising segmentations, with a mean Dice score of 0.7930 and 0.7432 for VS and cochlea respectively in the validation set.

Keywords: Unsupervised segmentation · Domain adaptation · Vestibular schwannoma · Cochlea · Deep learning.

1 Introduction

Vestibular schwannoma (VS) is a benign tumor of the human hearing system. For better understanding the disease progression, quantitative analysis of VS and cochlea from magnetic resonance images (MRIs) is important. Recently, deep learning frameworks have been dominating the medical segmentation field [5, 6, 1, 2, 4] with state-of-the-art performance. However, supervised learning methods may lack domain generalizability or ability to deal with MRIs from different domains, and such methods require a high level of consistency between training and testing data.

Furthermore, in the medical image analysis area, lack of human delineations in one or multiple domains is a common issue, which is problematic for supervised learning. Domain adaptation (DA) is a solution for increasing generalizability of deep learning models to deal with new data from different domains.

In this work, we propose an unsupervised cross-modality domain adaptation framework for segmenting the VS and cochlea. Our framework contains 2 parts: synthesis and segmentation. For synthesis, we apply a CycleGAN [7] to perform unpaired image translating between ceT1-weighted and T2-weighted MRIs. For segmentation, we use the generated T2-weighted MRIs as input and train an ensemble of models with various data augmentations, each of which yields the segmentations of VS and cochlea. We fuse those segmentations to form the final segmentation.

2   Methods and material

2.1   Dataset

The cross-modality domain adaptation for medical image segmentation challenge dataset[1] contains 2 different MRI modalities: contrast-enhanced T1-w with an in-plane resolution of 0.4×0.4mm and slice thickness between $1-1.5$mm, and high-resolution T2-w with an in-plane resolution of 0.5×0.5mm and slice thickness between $1-1.5$mm. ceT1-weighted imaging was performed with an MPRAGE sequence and T2-weighted imaging with 3D CISS or FIESTA sequence. The training set contains 105 ceT1-weighted and 105 T2-weighted MRIs, and the validation set contains 32 T2-weighted MRIs. Associated with all images in each modality are expert-acquired, manual VS and cochlea labels. More detailed information about this dataset can be found in [2].

2.2   Overall framework

Fig. 1 displays our proposed framework. The 3D ceT1w image, after undergoing pre-processing, is fed to a generator to achieve image-to-image translation (ceT1w to T2w). Data augmentation strategies are applied on the generated T2w to increase the model robustness. We send the augmented data into four different models: a 2.5D model with two different augmentation schemes, a 3D model with attention module, and a 3D model without attention module. We finally employ a union operation to fuse the outputs of each model to form the final segmentation.

2.3   Preprocessing and postprocessing

Our pre-processing pipeline contains 5 steps: (1) non-local mean filter denoising, (2) image alignment with MNI template by rigid registration, (3) bias field

---
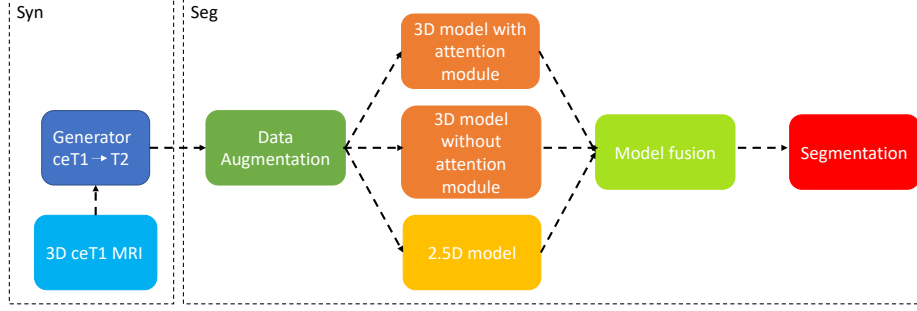
[1] https://crossmoda.grand-challenge.org/CrossMoDA/
[2] https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=70229053

Fig. 1. Proposed overall framework. Our framework contains 2 parts: synthesis (Syn) and segmentation (Seg). We use a CycleGAN model as generator in the synthesis part. Results from three different models with different augmentations are fused to obtain the final segmentation.

correction, (4) image cropping based on region of interest, and (5) intensity normalization. We extract the largest connected component for VS in postprocessing.

2.4   Data augmentation

Data augmentation is widely used in medical image segmentation to help minimize the gap between datasets/domains, producing more robust segmentations. Here, we design an 'online' augmentation strategy during training and randomly apply data transformations to input images. These transformations are categorized in 3 different groups:

- **Spatial augmentation.** 3 types of random spatial augmentation are used: affine transformation with angle range of $[-10°, 10°]$, and scale factor from 0.9 to 1.2; elastic deformation with control points $= 7$, max displacement $= 6$; and a combination of affine and elastic deformation. The same spatial augmentations and parameters are applied to both MRIs and the labels.
- **Image appearance augmentation.** To minimize the different image appearance between MRIs from different sites and scanners, we randomly apply multi-channel Contrast Limited Adaptive Equalization (mCLAHE) and gamma correction with γ from 0.5 to 2 to adjust image contrast.
- **Image quality augmentation.** In this context, image quality refers to resolution and noise level. We randomly blur the image using Gaussian kernel with σ_{blur} from 0.5 to 1.5, we add Gaussian noise with $\sigma_{noise} = 0.01$, and we sharpen the image by $I_s = I_b + (I_b - I_{bb}) \times \alpha$, where I_s is sharpened image, I_b is the image blurred with a Gaussian kernel (σ_{blur}), and I_{bb} is the image blurred twice with the same Gaussian kernel. In our case, we set $\alpha = 10, \sigma_{blur} = 1.5$.

2.5   2.5D model and its architecture

We applied a 2.5D CNN to alleviate the impact of the anisotropic image resolution. Network architecture details can be found in [6]. This 2.5D network contains both 2D and 3D convolutions to capture both in-slice and global information. Adapted from U-Net [3], the 2.5D model contains an encoder and a decoder. 2D convolutions are used in the first 2 levels and 3D convolutions for the remainder. The max-pooling and deconvolution operations connect the features between levels. Batch normalization and parametric rectified linear unit (pReLU) are used in the network architecture. An attention module is applied to assist in segmenting the small region of interest (ROI). A second 2.5D model with identical architecture was also used in the ensemble, with the only difference being in the gamma correction parameter in the augmentation stage (with $\gamma \in [0.5, 1.5]$ rather than $[0.5, 2]$).

2.6   3D model and its architecture

We used a fully convolutional neural network for our 3D models. The network architecture details can be found in [2]. Similar to a 3D U-Net, it consists of an encoder and a decoder. 3D max-pooling and 3D nearest neighbor upsampling are used in both the encoder and decoder. We further reinforce the output by adding the feature maps at the end of each level. In one of our two 3D models, we employ an attention module in the skip connections to emphasize the small ROI and preserve the information from encoder to decoder. Note that the two 3D models are identical except the attention module.

2.7   Implementation details

The Adam optimizer was used with L2 penalty of 0.00001, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and an initial learning rate of 0.0002 for the CycleGAN generators [7] and 0.001 for the segmentation models. For the generators, the learning rate was left at the initial value for the first 100 epochs, and then dropped to 0 in following 100 epochs. For segmentation models, learning rate was decayed by a factor of 0.5 every 50 epochs. We evaluated our generator and segmentation models every epoch and selected the optimal based on image quality by visual quality control and Dice score. We defined our loss function as $1 - \text{mean(Dice)}$ from multiple labels, with equal weight ($w_{FG} = 1$) for all foreground labels and decayed weight ($w_{BG} = 0.1$) for the background. The training, which has a batch size of 2, was conducted on NVIDIA GPUs and implemented using PyTorch.
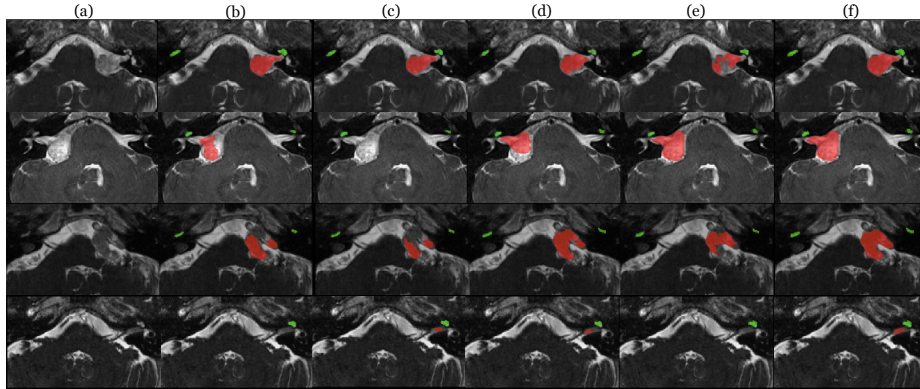
3   Results

Tab. 1 displays the quantitative results of our method in the validation phase, reported as mean(stdev). Representative qualitative results can be viewed in Fig. 2. In the first row of Fig. 2, we observe that our 2.5D baseline method

Table 1. Quantitative results in validation phase. The Dice score and average symmetric surface distance (ASSD) reported as mean(stdev).

	VS	Cochlea
Dice	0.7930(0.1515)	0.7432(0.0369)
ASSD	0.6343(0.3587)	0.2939(0.0626)

produces an accurate segmentation. Varying the augmentation parameters allows the 2.5D model to perform better in some challenging cases (fourth row of Fig. 2-(c)). However, in some cases, both 2.5D models under-segments VS, which can be seen in the second and third rows of Fig. 2-(b,c). In such cases, 3D models compensate for this tendency of the 2.5D models. Although the attention module has good ability to capture small ROIs, which can be observed in the second row of Fig. 2-(e), it may also lead to over-fitting (first and third rows of Fig. 2-(e)). Thus, we also include a 3D model without attention module (first and third rows of Fig. 2-(d)) in the model ensemble for best results. The model fusion balances the strengths of the individual models and is able to produce consistently good segmentations in a variety of images (Fig. 2-(f)).

Fig. 2. Segmentation results from 4 different subjects. Red, VS, green, cochlea. (a) Input image. (b) 2.5D model with $\gamma \in [0.5, 2]$. (c) 2.5D model with $\gamma \in [0.5, 1.5]$. (d) 3D model without attention. (e) 3D model with attention. (f) Final segmentation. The final model fusion step produces consistently good segmentations.

4   Conclusion

In this work, we proposed an unsupervised cross-modality domain adaptation framework for VS and cochlear segmentation. In the validation stage, our method shows promising results.

References

1. Dorent, R., Joutard, S., Shapey, J., Bisdas, S., Kitchen, N., Bradford, R., Saeed, S., Modat, M., Ourselin, S., Vercauteren, T.: Scribble-based domain adaptation via co-segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 479–489. Springer (2020)
2. Li, H., Zhang, H., Johnson, H., Long, J.D., Paulsen, J.S., Oguz, I.: Mri subcortical segmentation in neurodegeneration with cascaded 3d cnns. In: Medical Imaging 2021: Image Processing. vol. 11596, p. 115960W. International Society for Optics and Photonics (2021)
3. 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. Springer (2015)
4. Shapey, J., Kujawa, A., Dorent, R., Wang, G., Dimitriadis, A., Grishchuk, D., Paddick, I., Kitchen, N., Bradford, R., Saeed, S.R., Bisdas, S., Ourselin, S., Vercauteren, T.: Segmentation of vestibular schwannoma from mri — an open annotated dataset and baseline algorithm. Scientific Data (2021), in press. Preprint available at https://doi.org/10.1101/2021.08.04.21261588 medRXiv:10.1101/2021.08.04.21261588
5. 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)
6. Wang, G., Shapey, J., Li, W., Dorent, R., Demitriadis, A., Bisdas, S., Paddick, I., Bradford, R., Zhang, S., Ourselin, S., 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. pp. 264–272. Springer (2019)
7. 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)