Cross-Modality Domain Adaptation for Vestibular Schwannoma and Cochlea Segmentation from High-Resolution T2 MRI (Epione-Liryc team)

Buntheng Ly[1], Victoriya Kashtanova[1], Yingyu Yang[1], Aurélien Maillot[2,3], Marta Nuñez-Garcia[2,3], and Maxime Sermesant[1,2]

[1] Inria, Université Côte d'Azur, Epione team, Sophia Antipolis, France
[2] IHU LIRYC, Electrophysiology and Heart Modeling Institute, Pessac, France
[3] Université de Bordeaux, Bordeaux, France

Abstract. Segmentation of vestibular schwannoma (VS) from MRI is commonly performed using contrast-enhanced T1 (ceT1) scans. However, high-resolution T2 (hrT2) imaging has been recently shown as a reliable, safer, and lower-cost alternative to ceT1 for the same task. In the context of the Cross-Modality Domain Adaptation (CrossMoDa) challenge 2021, we proposed an automatic pipeline for the segmentation of the VS and the cochlea from hrT2 MRI. The provided training set contained unpaired annotated ceT1 and non-annotated hrT2 scans. In our approach, we first spatially normalised the input images to MNI space which allowed us to better identify the global region of interest using the ground truth labels. Then, cross-modality domain adaptation was performed by training and using a CycleGAN model to translate the ceT1 to *fake* hrT2. The segmentation model was built using the 3D U-Net architecture and was trained fully supervised using the original and generated dataset. To further refine the masks, clustering methods such as k-Means, Mean-Shift, and Dense Conditional Random Field were applied in the original spatial domain. Our approach was tested using the validation set provided by the organizers (32 hrT2 scans) obtaining a mean Dice score of 0.7838 ± 0.0989 (0.7932 ± 0.1772 and 0.7745 ± 0.0416 for VS and cochlea, respectively).

Keywords: Domain Adaptation · CycleGAN · UNet · Image Segmentation.

1 Methods

All images were firstly spatially normalized to MNI space and a bounding box containing the VS and cochlea using the corresponding label images was computed (section 1.1). A CycleGAN model was trained to convert the spatially normalised ceT1 images to *fake* spatially normalised hrT2 images (section 1.2). The generated *fake* hrT2 images, along with the corresponding ceT1, were then used to train a segmentation model based on the 3D U-Net in a fully supervised manner (section 1.3). Finally, the output segmentations were converted back to their original space where they were further refined (section 1.4). All methods were implemented in Python.

1.1 Normalisation

The SPM12[4] Matlab tool and a Python interface (nipype[5]) were used to map the ceT1 and corresponding masks, and the hrT2 images into the standard MNI space. By applying the Segment function (default setting), the forward and inverse deformation fields were obtained. The ceT1 and hrT2 images were resampled to an isotropic 0.5 mm spacing resolution using the forward deformation field. Linear interpolation was used for the images, and nearest neighbor interpolation for the masks. A general bounding box of the two targets (the two cochlea and VS) using the addition of all transformed label images was then computed.

1.2 Image Translation

A CycleGAN model [4] was used to create a training database of labeled hrT2 scans. This self-supervised method is able to perform image-to-image translation in the case of unpaired datasets.

Training database Since only 105 items of 3D-scans were available per dataset for training, we decided to use the set of 2D-slices to train the CycleGAN. From each normalised scan 218 slices were extracted along the z-axis, obtaining 22890 2D images per dataset in total. Each 2D slice was resized to 256×256 pixels using SimpleITK[6] resize function with linear interpolation, and the image intensity was normalised to $[0 - 1]$.

CycleGAN model configurations We used the standard CycleGAN parameters described in the paper [4] with 200 epochs of training, except for the number of filters in the generator (ResNet) and discriminator networks that were empirically set to 16.

Pair-Loss To improve the performance of CycleGAN we added a supervised regularisation technique to control the training process, called the Pair-Loss. Using the cross-entropy metric as base, we semi-automatically selected 8 pairs of closest 2D-slice images from 2 training databases which we used as image-pairs for the Pair-Loss. The Pair-Loss was created similarly to Cycle consistency loss from the original CycleGAN paper [4] and was calculated as a sum of MSE-losses between generated *fake* images and the original ones for each image pair,

[4] https://www.fil.ion.ucl.ac.uk/spm/software/spm12/

[5] https://nipype.readthedocs.io/en/latest/

[6] https://simpleitk.org/

according to the formula:

$$\sum_{p \in Pairs} (MSE(G_{T1->T2}(T1_p), T2_p) + MSE(G_{T2->T1}(T2_p), T1_p))$$

This additional metric significantly boosted quality of the generated *fake* images.

Using the trained with Pair-Loss CycleGAN model we generated *fake* spatially normalised hrT2 dataset with associated segmentation labels from the spatially normalised ceT1 dataset.

1.3 Image segmentation

A U-Net model was used as the segmentation network [2]. The network consists of three levels of encoding and decoding and one deepest level convolution unit. Each unit was built as two convolutional layers each one followed by a rectifier activation and instance normalisation layer. The numbers of convolutional filters were set to 16, 32, 64 and 128 from the first to the deepest level. We used 3D max pooling and upsampling layer to decrease and increase the feature resolutions and the concatenation layer to link the skip connection features.

To train the model, the normalised ceT1 dataset was splitted into training, validation and testing dataset using $8:1:1$ ratio. The images were cropped using the bounding box defined in section 1.1, resized to $144 \times 80 \times 80$ and normalised the intensity to $[0-1]$ to be used as input to the segmentation network.

For training, we used both the original ceT1 images and their *fake* hrT2 images while only the *fake* hrT2 images were used for validation and testing. To further increase the number of training data, we applied random image augmentations on the input image. The augmentation was done during the training with the 80% probability rate and could be any combination of rotation, Gaussian noise and midsagittal flip.

We trained two separate models for the segmentation of the cochlea and the VS. The network was trained using the sum of Dice and the distance loss. The distance loss is calculated using equation 1, where y is the segmentation output and m is the euclidean distance map of the ground truth mask.

$$L_{dist} = \frac{1}{N} \sum^{N} y * m \tag{1}$$

The segmentation networks were trained using Adam optimiser with the starting learning rate at $1e-4$. The learning rate was set to reduce by half after 5 epochs of no validation improvement, and the training was set to stop after 20 epochs without validation improvement. We applied evaluation on the testing dataset at each epoch, and the final weights were selected based on the best evaluation Dice score.

1.4 Segmentation refinement

The segmentation outputs were transformed back to their original space using the inverse deformation fields computed as explained in section 1.1. They were further post-processed as follows.

For the cochlea, the preliminary mask was slightly dilated 2 voxels in lateral and superior-inferior (X and Y) axes. Then, k-Means clustering using the intensity values of the dilated cochlea was applied. The number of clusters was set to 2 and the cluster with highest mean intensity was kept.

For the VS mask, we applied 1mm dilation in lateral and superior-inferior (X and Y) axes. Then, k-Means and Mean-Shift clustering were applied using both the intensity and the coordinates of the dilated VS. The number of clusters for k-Means was set to 3 and the cluster with medium mean intensity was kept (i.e. the darkest and brightest clusters were removed). Related to Mean-Shift, the biggest cluster was kept. Majority voting was applied using the preliminary VS mask, and the k-Means and Mean-Shift derived masks. Voxels with two or more votes were selected. Morphological operations such as binary opening and closing were used to refine the masks. Finally, we used Dense Conditional Random Field (CRF), frame by frame, to further improve the segmentation result. The ROI is usually a rectangle that covers frame VS mask, with 20-pixel margin in each direction. The post-processed tumour mask (hard labelling) was used to define the unitary potentials. Both positional and colour-dependant features were applied to build pairwise potentials. For the colour vector in pairwise potential, along with the hrT2 intensity we added a second channel: a Gaussian Mixture Clustering Map with 4 mixture components. The parameters of the model were manually set to $\theta_\alpha = 7$ and $\theta_\beta = 0.9$ (see function (3) in [1] based on visual examination of the target validation samples. The inference output was obtained by mean field approximation (with iteration number set to 200). The final VS tumour segmentation was the Dense CRF inference output.

2 Results

105 ceT1 scans with corresponding VS and cochleas segmentations, and unpaired 105 hrT2 scans were made available by the organizers for training. More details about this dataset can be found in [3]. 32 hrT2 scans were released during the evaluation period. Using our pipeline, the mean Dice score on the validation dataset was 0.7838 ± 0.0989. For the VS, the mean Dice was 0.7932 ± 0.1772 and the mean Average Symmetric Surface Distance (ASSD) was 1.4986 ± 4.9768 mm. For the cochlea, the mean Dice was 0.7745 ± 0.0416 and the mean ASSD was 0.3286 ± 0.5316 mm.

References

1. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected crfs with gaussian edge potentials. In: Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems. vol. 24. Curran Associates, Inc. (2011)
2. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **9351**, 234–241 (2015)

3. 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/10.1101/2021.08.04.21261588medRXiv:10.1101/2021.08.04.21261588
4. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Computer Vision (ICCV), 2017 IEEE International Conference on (2017)