Tumor blending augmentation using one-shot generative learning for crossmodal MRI segmentation

Guillaume Sallé, Pierre-Henri Conze, Julien Bert, Nicolas Boussion, Ulrike Schick, Dimitris Visvikis and Vincent Jaouen

UMR 1101 Inserm LaTIM, UBO, IMT Atlantique, Brest, France
guillaume.salle@univ-brest.fr

Abstract. In the context of the CrossMoDa 2022 challenge, we propose a new domain adaptation technique for vestibular schwanomma (VS) and cochlear segmentation. We use a CycleGAN translation model combined to a new data augmentation method based on a generative network trained on a single image. The method, called tumor blending augmentation (TBA), allows to realistically diversify the appearance of target regions of interest in the training set while leaving the rest of the image unchanged, exposing a downstream segmentation network to a wider range of appearances and thus improving its generalization ability at test time. Our solution ranked first at the end of the validation stage of the challenge, with average Dice scores of 0.8682 ± 0.0601 for VS and 0.8506 ± 0.0294 for cochlea.

Keywords: Unsupervised domain adaptation · Vestibular schwannoma · Cochlea · Cross-modal segmentation · One-shot Learning

1 Introduction

In most current clinical routine for radiation therapy treatment planning of vestibular schwanomma (VS), the tumor and the organ at risk cochlea are segmented on contrast-enhanced T1-weighted images (ceT1) [1]. Recently, dedicated MRI sequences such as high resolution T2 images (hrT2) have raised interest in order to reduce global costs and the use of gadolinium contrasting agents [2],[3]. Due to the cost of producing annotations in other modalities, unsupervised domain adaptation models that reuse previously labelled images could be of great significance [4].

Generative Adversarial Networks (GANs) are often considered for artificial data augmentation in machine learning-based models to address data scarcity, a recurring issue in medical imaging [5]. However, most GAN-based methods rely on large training sets in order to expose both the generator and the discriminator to sufficient number of examples. Departing from these usual requirements, the one-shot 2D GAN SinGAN model [6] has recently emerged as a new paradigm for deep generative learning using multiple adversarial generators in a cascaded

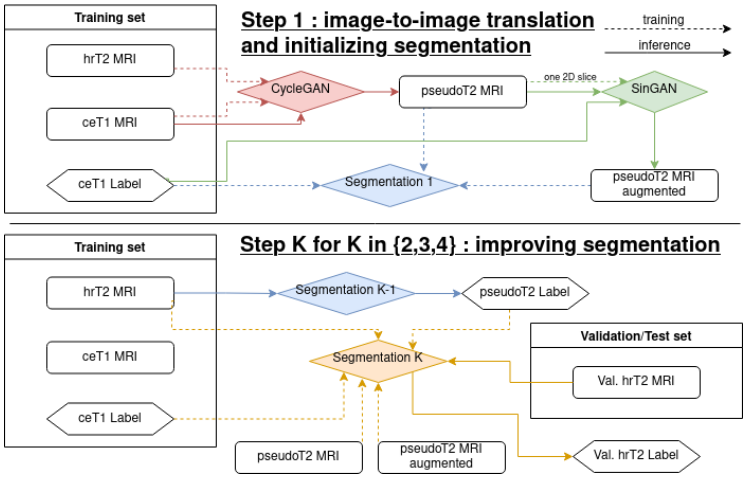


Fig. 1: General workflow of the proposed approach. VS and cochlea are always segmented separatly. The last segmentation is trained with hrT2 and pseudoT2 labels only. For cochlea, K varies from 2 to 3 only.

multi-scale architecture. SinGAN enables to learn the rich complexity of natural images using as little as one training image [7],[8].

In this work, we propose a new unsupervised cross-modal domain adaptation pipeline for VS and cochlear segmentation. In the context of the CrossMoDA 2022 segmentation challenge, we propose to leverage a SinGAN model trained on a single axial slice to realistically blend synthetically altered versions of the regions of interest (ROI) for efficient data augmentation. Such a technique makes the training more robust by increasing the model's generalization ability. We show that this approach improves segmentation quality, reaching state of the art cross-modal segmentation performance.

2 Method

Fig. 1 shows the general workflow of the proposed pipeline. We first train a 2D CycleGAN image-to-image translation network [9] to translate ceT1 images into pseudo hrT2 scans. Pseudo hrT2 scans are then used with ceT1 masks to train a segmentation network based on pseudo labels.

CycleGAN models perform a global image mapping that may however make small regions of interest (ROI) such as small VS or cochlea, less visible [10]. This is indeed what we observed during translation for both ROI, leading to unrealistic appearance. To increase the realism of the hrT2 set and therefore improve the generalization ability of the network during the downstream segmentation task, we employ a tailored data augmentation based on a 2D one-shot multi-stage generative SinGAN model [6] to generate more realistic variations of appearance of both VS and cochlea. Given a single image, SinGAN learns the image distribution at $N+1$ different scales using $N+1$ scale-specific generators trained

successively in a coarse-to-fine fashion. The general principles of the SinGAN model are summarized in Fig. 2.

After training on a 2D axial slice from a real hrT2 image, $N+1$ generators are available for $N+1$ different scales. First, each ROI to be augmented is naively brightened or darkened using a multiplicative intensity scaling factor λ, as we observed ROIs to present variable contrast with background in the training set. We then fix a generator scale n_0 and apply successively all higher level generators $n \geq n_0$ on each 2D slice were the ROI is visible. Finally, the 2D images are tiled back to reconstitute a 3D volume.

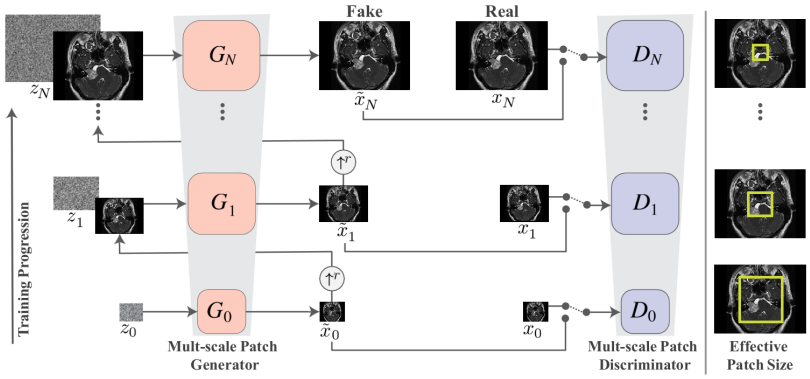
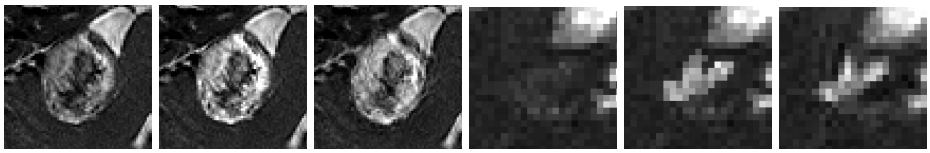


Fig. 2: SinGAN training pipeline. The image is initially downsampled to a very low scale, numbered 0. At each scale n, the generator G_n learns to synthesize realistic image patches. It gradually improves the realism of the output from the previous scale $n-1$, while the adversarial discriminator D_n learns to distinguish real and generated samples. After training the scale n, the result is upsampled by factor r up to the next scale. Modified from [6].

We perform the segmentation of the cochlea and of the VS independently. A major reason for this is that ground truth VS segmentation masks were generally obtained from ceT1 images, while the cochlea was most often labelled on the hrT2 image [11]. The cochlea is very difficult to visualize on ceT1 images, which leads us to not rely on the generation of pseudo hrT2 from ceT1 for pseudo labelling of the cochlea, contrary to VS. For the cochlea, we thus only consider augmented pseudo hrT2 while we use both original and augmented pseudo hrT2 for VS. After training the first segmentation networks for each structure, we predicted masks on the real hrT2 images to generate pseudo hrT2 labels. To further improve the quality of the pseudo labels, we repeat this self training stage iteratively two more times [12].

From our first experiments we noticed that a certain number of VS from center ETZ were badly predicted by the downstream segmentation model. These

tumors were generally heterogeneous, large, and showing hypersignal regions that the network was not exposed sufficiently to. To tackle this issue, all VS from ETZ of volumes larger than 2340 mm^3 with standard variation higher than 0.09 (6500 voxels for 29 images in total) were augmented with tumor blending augmentation using intensity scaling factors λ of 0.7, 1.2 and 1.5. We selected these values to augment around a third of center ETZ. We also augmented images with VS volumes less than 288 mm^3 (800 voxels ; 19 images in total) by using λ of 0.6, 0.8 and 1.2 to increase the proportion of weakly appearing tumors. Due to the weak appearance of the cochlea on ceT1 images, and even more on pseudo hrT2s, tumor blending intensity scaling factors λ of 2, 3 and 4 were considered for this organ. To also facilitate the first cochlear segmentation stages, true hrT2 were clipped to the 95th intensity percentile as the organ generally appears as hypersignal in hrT2 images.



(a) (b) (c) (d) (e) (f)

Fig. 3: Augmentation results on training pseudoT2 images. (a,d) original pseudoT2, (b,e) naively rescaled VS or cochlea ($mask \times 4.0$ for cochlea, $mask \times 1.5$ for VS), (c,f) augmented pseudoT2.

Due to the time constraints of the challenge and reduced frequency of optimization on the validation set, the λ values were selected based on non exhaustive quality judgment to yield a diverse training set showing smaller domain shift with respect to the validation set. The proposed method was robust to variations of these values, provided their choice sufficiently overlapped with the range of ROI contrasts seen at test time. More optimal parameters could however naturally be determined automatically through grid search optimization.

3 Experiments and results

All scans were resampled to a voxel spacing of $0.6 \times 0.6 \times 1.0$ mm^3. We then cropped a $256 \times 256 \times Z$ volume (with Z the number of axial slices) by computing the x and y average location of voxels higher than the 75th percentile to identify the center of the brain, as proposed in [13] for last year's challenge [14]. The selected SinGAN scales were 13 and 15 for VS (2 augmentations with the same images) and 13 only for cochlea. SinGAN scaling factor is 0.85, which makes $N = 17$. Fig. 3 shows a selection of tumor blending augmentation result.

The downstream segmentation task was performed by a 5-fold ensemble 3D full resolution nnUNet [15] model for 500 epochs. The last segmentation model

was trained for 1000 epochs using real hrT2 with the last best pseudo labels only. This step showed to slightly further improve performance. The largest connected components was kept for every VS model.

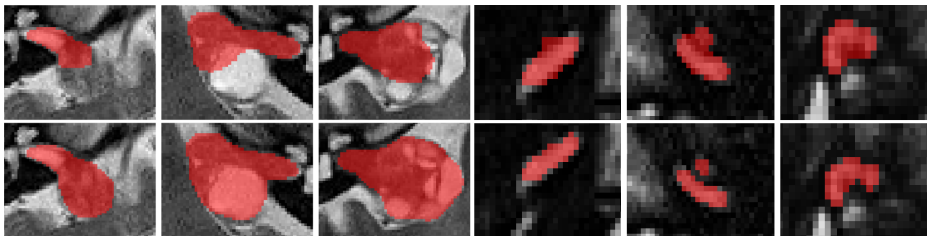


Fig. 4: Segmentation results from validation set at step 2 without (first row) and with (second row) our data augmentation technique.

Before the last segmentation network for each ROI, we resampled images to a finer voxel spacing of $0.4 \times 0.4 \times 1.0$ mm and extracted the same crop to refine the segmentation masks. This step increased the level of detail of the predicted labels, as it is closer to original spacing. Our CycleGAN was trained with all axial slices from London center only. The images were also deconvolved with an iterative Van Cittert deconvolution algorithm [16], as this step made contours sharper and facilitated the later segmentation step. For VS, we used a Gaussian point spread function (PSF) of scale $1 \times 1 \times 2.5$ mm^3 for 15 iterations. At the last inference for cochlea, input images were also sharpened using a Van Cittert algorithm using softer parameters $0.4 \times 0.4 \times 1.5$ mm^3 for 15 iterations.

	DICE score	ASSD
VS	0.8682 ± 0.0601	0.4302 ± 0.1780
Cochlea	0.8506 ± 0.0294	0.1892 ± 0.1457

Table 1: Best DICE score and ASSD obtained on validation set.

Representative segmentation results are shown in Fig. 4, illustrating the advantage of our TBA stage for improving segmentation quality over using only conventional synthetic data augmentation as in nnUNet. Our solution ranked first on the validation set, with average Dice scores reported in table 1.

4 Conclusion

In this study, we proposed a new data augmentation technique based on a generative adversarial network trained on a single image to realistically blend objects in medical images for improved generalization of segmentation algorithms. The proposed approach is key to a global workflow for cross-modal VS and cochlear segmentation in the CrossMoDA 2022 challenge, where we obtained the first rank during the validation phase.

References

1. Elizabeth A Vokurka et al., "Using bayesian tissue classification to improve the accuracy of vestibular schwannoma volume and growth measurement," *American journal of neuroradiology*, vol. 23, no. 3, pp. 459–467, 2002.
2. Daniel H Coelho et al., "Mri surveillance of vestibular schwannomas without contrast enhancement: clinical and economic evaluation," *The Laryngoscope*, vol. 128, no. 1, pp. 202–209, 2018.
3. Guotai Wang et al., "Automatic segmentation of vestibular schwannoma from t2-weighted mri by deep spatial attention with hardness-weighted loss," in *MICCAI*. Springer, 2019, pp. 264–272.
4. Reuben Dorent et al., "Scribble-based domain adaptation via co-segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 479–489.
5. Junzhao Liang and Junying Chen, "Data augmentation of thyroid ultrasound images using generative adversarial network," in *2021 IEEE International Ultrasonics Symposium (IUS)*, 2021, pp. 1–4.
6. Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli, "Singan: Learning a generative model from a single natural image," in *ICCV*, 2019, pp. 4570–4580.
7. Vajira Thambawita et al., "Singan-seg: Synthetic training data generation for medical image segmentation," *PloS one*, vol. 17, no. 5, 2022.
8. Guillaume Sallé, Pierre-Henri Conze, Nicolas Boussion, Julien Bert, Dimitris Visvikis, and Vincent Jaouen, "Fake tumor insertion using one-shot generative learning for a cross-modal image segmentation," *IEEE NSS-MIC*, 2021.
9. Jun-Yan Zhu et al., "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks," in *ICCV*, 2017.
10. Joseph Paul Cohen, M. Luck, and Sina Honari, "How to Cure Cancer (in images) with Unpaired Image Translation," in *MIDL 2018*.
11. Jonathan Shapey et al., "Segmentation of vestibular schwannoma from mri, an open annotated dataset and baseline algorithm," *Scientific Data*, 2021.
12. Hyungseob Shin et al., "Cosmos: Cross-modality unsupervised domain adaptation for 3d medical image segmentation based on target-aware domain translation and iterative self-training," *arXiv:2203.16557*, 2022.
13. Jae Won Choi, "Using out-of-the-box frameworks for unpaired image translation and image segmentation for the crossmoda challenge," *arXiv e-prints*, 2021.
14. Reuben Dorent, Aaron Kujawa, Marina Ivory, Spyridon Bakas, Nicola Rieke, Samuel Joutard, Ben Glocker, Jorge Cardoso, Marc Modat, Kayhan Batmanghelich, et al., "Crossmoda 2021 challenge: Benchmark of cross-modality domain adaptation techniques for vestibular schwnannoma and cochlea segmentation," *arXiv preprint arXiv:2201.02831*, 2022.
15. Fabian Isensee et al., "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, 2021.
16. PH Van Cittert, "Zum einfluß der spaltbreite auf die intensitätsverteilung in spektrallinien," *Zeitschrift für Physik*, pp. 547–563, 1930.