nn-UNet Training on CycleGAN-translated images for cross-modal domain adaptation in biomedical imaging

Smriti Joshi, Richard Osuala, Carlos Martın-Isla, Victor M. Campello, Carla Sendra-Balcells, Karim Lekadir, and Sergio Escalera

Artificial Intelligence in Medicine Lab (BCN-AIM), Faculty of Mathematics and Computer Science, Universitat de Barcelona, Spain

Abstract. In recent years, deep learning models have considerably advanced the performance of segmentation tasks on Brain Magnetic Resonance Imaging (MRI). However, these models show a considerable performance drop when they are evaluated on unseen data from a different distribution. Since annotation is often a hard and costly task requiring expert supervision, it is necessary to develop ways in which existing models can be adapted to the unseen domains without any additional labelled information. In this work, we explore one such technique which extends the CycleGAN architecture to generate label-preserving data in the target domain. The synthetic target domain data is used to train the nn-UNet framework for the task of multi-label segmentation. The experiments are conducted and evaluated on the dataset [1] provided in the 'Cross-Modality Domain Adaptation for Medical Image Segmentation' challenge for segmentation of vestibular schwannoma (VS) tumour and cochlea on contrast enhanced (ceT1) and high resolution (hrT2) MRI scans. In the proposed approach, our model obtains dice scores (DSC) 0.72 and 0.49 for tumour and cochlea respectively on the validation set of the dataset. This indicates the applicability of the proposed technique to real-world problems where data may be obtained by different acquisition protocols as in [1] where hrT2 images are more reliable, safer, and lower-cost alternative to ceT1.

Keywords: Domain Adaptation · Vestibular schwannoma (VS) · Deep Learning · nn-UNet · CycleGAN

1 Introduction

Deep learning techniques have achieved immense success under the assumption that the training data and test data come from the same distribution. However, this assumption frequently does not hold true in real world data due to differences in acquisition conditions and techniques. In recent works, unsupervised domain adaptation has proven to be an important tool to traverse this domain gap between source data and un-annotated target data. These techniques reduce the need of expensive labelling in target domain without compromising

the performance of the model. For example, in semantic segmentation of street view for autonomous vehicles, one real image of Cityscapes[15] dataset takes 1.5 hours to annotate. Unsupervised domain adaptation (UDA) techniques enable researchers to train models on synthetic data (for which labels can be easily generated) and adapt it to real data. In the field of medical image analysis, annotating images is not only expensive and time-consuming but it also requires tedious and labor-intensive participation of physicians, radiologists and other experts. Further, domain gaps are observed in medical data due to images obtained from different clinical centres, imaging conditions, scanner vendors [14] and the modality of the data. Since medical field has low error tolerance, it is essential to devise techniques that can perform well even in the presence of high domain gap. In this work, we propose one such approach for segmentation of vestibular schwannoma (VS) tumour and cochlea on unannotated high-resolution T2 (hrT2) brain MRI scan given fully annotated contrast-enhanced T1 (ceT1) data.

Over the years, many techniques based on divergence minimization [5][6], adversarial learning [7], normalization [8] and domain disentanglement [9] have been proposed for domain adaptation/generalization. More recent techniques based on self-supervision also have performed well on datasets with smaller domain gaps. In this work, we use a domain mapping technique extending the CycleGAN architecture to map brain MRI scans from source domain (ceT1) to target domain (hrT2) and train a segmentation network with available source domain annotations.

2 Method

The method adopted for domain adaptation in this work has two main steps:

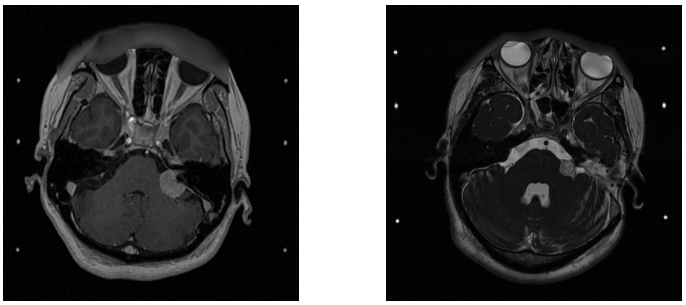
1. Generation of synthetic images in hrT2 domain using the CycleGAN architecture. [2]
2. Training segmentation model using nn-UNet [3] framework with generated images from various stages of CycleGAN training.

The following sections discuss this pipeline in more detail. Figure 2 presents the main steps of the pipeline.

2.1 Dataset and Evaluation Metrics

The dataset[1] used for experiments and evaluation of this work consists of unpaired MRI scans with ceT1 and hrT2 modalities. For the ceT1 domain, tumour and cochlea are annotated as 1 and 2 respectively while for the hrT2 domain, no label information is available. The ceT1 images have pixel dimension of 512 x 512 with a resolution of 0.4 x 0.4 mm while hrT2 images are of pixel dimensions 384 x 384 or 448 x 448 with higher resolution of 0.5 x 0.5 mm. Further, in hrT2 images, the number of slices along the z axis is either 20, 40 or 80. Figure 1 shows the resized axial slices of ceT1 and hrT2 images for visual demonstration of the domain gap between the two modalities. Evaluation of the experiments

is conducted on the validation set of this dataset with Dice score (DSC) and Average Symmetric Surface Distance (ASSD) as performance metrics.



(a) ceT1 (b) hrT2

Fig. 1: Axial slices of MRI modalities present in the dataset showing VS tumour

2.2 Setup

For the development of the project, Python 3.8 was used with PyTorch 1.7.1 framework. The experiments were conducted on the BCN-AIM servers on Nvidia RTX 3090 GPU with 24 GB memory.

2.3 Preprocessing

Slice selection: To train the standard *CycleGAN* architecture, the 3D data is saved as 2D axial slices for each scan in grayscale format. Looking closely at the data, it can be seen that cochlea (and tumour) is usually located in the initial axial slices of the images. Therefore, to minimize the occurrence of irrelevant data, the range of slices containing labelled information is calculated and only the slices inside this range ([10, 60] are used for training. These slices are also resized to a common size of 192 x 224. Further, to train the segmentation model using *nn-UNet*, these 2D slices in the ceT1 domain are mapped to the hrT2 domain using the trained CycleGAN and concatenated to form a 3D volume of size 192 x 224 x 51.

2.4 Generating synthetic images

The generation of synthetic images is conducted using 2D CycleGAN[1] using the axial slices of MRI scans. Figure 3 shows the translation of images from ceT1 domain to hrT2 domain. It can be observed that tumour information is preserved during the translation.

[1] Code available at: `https://github.com/aitorzip/PyTorch-CycleGAN`

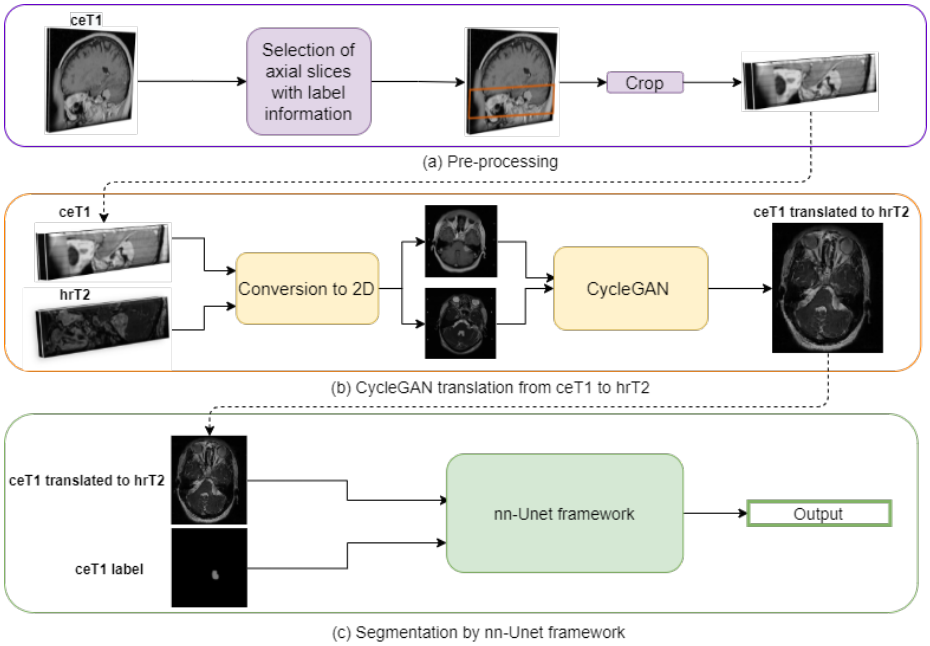


Fig. 2: Description of main steps of the pipeline. 3D images are presented from sagittal view to demonstrate the slice selection in axial direction. In 2D, slices presented are axial

Configuration: Before feeding the image to the network, the images are normalized to the range $[-1, 1]$. Adam optimizer with initial learning rate of 0.00002 is used to train the generators as well as the discriminators. For augmentation, basic torchvision transforms namely RandomHorizontalFlip (probability: 0.5) and RandomRotation (5°) are used. The remaining parameters are same as in the original configuration [2]. The model is trained for 100-150 epochs depending upon visual evaluation of translated images.

To make sure that the discriminator does not overpower the generator, it is only updated when its accuracy gets too low, that is, below 0.6.

2.5 Segmentation for hrT2 domain

Once the CycleGAN is trained, it is used to map the ceT1 data to hrT2 domain. To implement this, the individual slices are mapped and concatenated to get a volume of size 192 x 224 x 51. The labels corresponding to these translated images are corresponding slices of given ceT1 ground truth images. Further, multiple sets of images are translated using CycleGAN weights from different stages. One observation inspiring this approach was the difference in representation of tumours in generated images over the training epochs. Initial epochs of CycleGAN show brighter tumours similar to ceT1 while darker tumours are

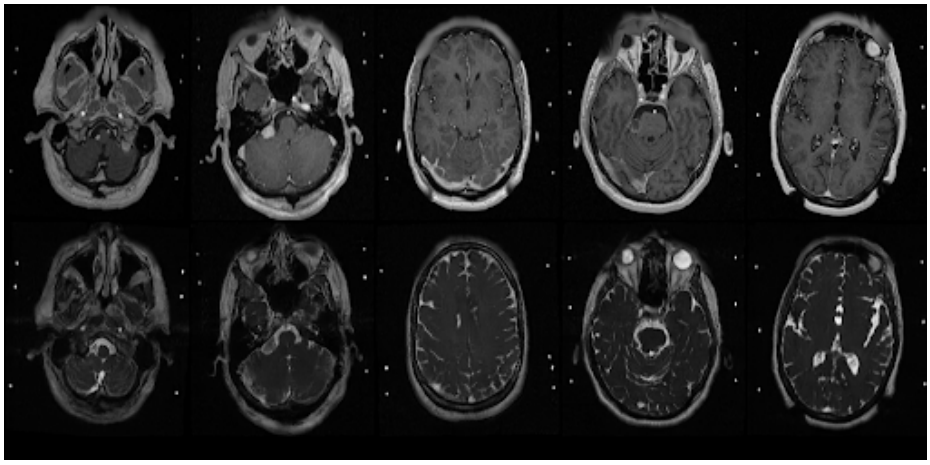


Fig. 3: The translation from ceT1 to hrT2 domain using CycleGAN for non-registered data. First row corresponds to the randomly generated axial slices corresponding to ceT1 domain while the second row shows corresponding translated synthetic images in hrT2 domain.

obtained in later epochs. Since both kinds of textures are present in hrT2 data, this approach was adopted to improve generalization of the network.

Our network is built upon the nn-UNet [3], a dynamic fully automatic segmentation framework for medical images leveraging the UNet, which is commonly used for segmentation tasks. For training this network, the nn-UNet configuration for 2D data is used. This is because that CycleGAN is trained on 2D images and information along the z-axis cannot be taken into consideration due to its anisotropic nature. This is also supported by the experimentation as discussed in the section 3. As a part of nn-UNet pre-processing, the data is cropped (region of non-zero values only), resampled and normalised. Loss function used for training the 2D nn-UNet is a combination of dice loss and cross entropy loss. Instead of using instance normalisation as in the original configuration, batch normalization is used as the authors of [4] demonstrated that it yields better performance on brain MRI segmentation tasks.

Augmentation: Data augmentation is an integral part of the nn-UNet framework. The following augmentations are used for training the framework in this work: Spatial transforms such as elastic deformation, rotation, scaling, random crop and intensity transformations like gaussian noise, gaussian blur, additive brightness, multiplicative brightness, contrast, simulated low resolution, gamma transform and mirroring. The file used for data augmentations can be accessed from data_augmentation_moreDA.py [2] file of nn-UNet framework.

[2] code available at: `https://github.com/MIC-DKFZ/nnUNet`

Postprocessing in nn-UNet: For the tumour labels, all predictions but the one with the largest region in the 3D volume are removed as it indicated better performance on internal validation set generated during five-fold cross validation training by nn-UNet framework.

Method	**Tumour**				**Cochlea**				**Total**	
	Dice		ASSD		Dice		ASSD		Dice	
	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.
3D nn-UNet	0.6072	0.2766	6.1097	8.1267	0.3212	0.0682	1.0882	2.5184	0.4642	0.1458
2D nn-UNet	0.6572	0.2131	2.3072	6.7850	0.4388	0.0980	1.0964	2.4962	0.5480	0.1224
+ pretrain	0.6999	0.1470	1.0188	0.5377	0.4460	0.0828	0.6602	0.2160	0.5730	0.0861
+ pretrain + MD	0.7120	0.1411	0.9643	0.5117	0.4779	0.0773	0.6024	0.1874	0.5950	0.0860
+ pretrain + MD + PL	**0.7291**	**0.1428**	**0.8969**	**0.4929**	**0.4944**	**0.0598**	**0.5591**	**0.1573**	**0.6117**	**0.0813**

Table 1: Evaluation metrics on validation set. MD refers to the case where more data generated from CycleGAN weights over epochs is added to the dataset, PL refers to pseudo-labelling.

3 Results

On the validation data, the metrics obtained with our approach with different configurations are summarised in Table 1. 2D nn-UNet performs better than 3D nn-UNet. This is intuitive as the translated data obtained from 2D CycleGAN which likely lack some 3D coherence across domains. Pretraining the network with labelled ceT1 data makes tumour recognition better contributing to an increase of ≈ 0.03 to the overall dice score. Further, adding more synthetic data from different CycleGAN weights to the training set pushes the performance even further. Addition of real target data to the training data with pseudo-labels predicted by the former configuration also pushes the performance by overall dice value of ≈ 0.02. Generally, the results demonstrate that tumour is identified better than cochlea. This is expected since cochlea has a smaller structure and is less pronounced compared to the tumour making it more difficult to translate from ceT1 to hrT2 domain.

4 Conclusion

Finally, the highest overall dice score obtained is 0.6117 with higher recognition of tumour (≈ 0.73) than cochlea (≈ 0.49). In the future work, we will aim at improving cochlea recognition by utilising information about its position in the brain as the anatomy can be expected consistent across all patients.

References

1. J. Shapey, A. Kujawa, R. Dorent, G. Wang, A. Dimitriadis, D. Grishchuk,I. Paddick, N. Kitchen, R. Bradford, S. R. Saeed, S. Bisdas, S. Ourselin,and T. 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.
2. Zhu, J.-Y., Park, T., Isola, P. & Efros, A. A. (2017). Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks.. CoRR, abs/1703.10593.
3. Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P. F., Kohl, S., Wasserthal, J., Koehler, G., Norajitra, T., Wirkert, S. J. & Maier-Hein, K. H. (2018). nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation.. CoRR, abs/1809.10486.
4. Isensee, F., Jaeger, P. F., Full, P. M., Vollmuth, P. & Maier-Hein, K. H. (2020). nnU-Net for Brain Tumor Segmentation.. CoRR, abs/2011.00848.
5. Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B. & Smola, A. J. (2006). A Kernel Method for the Two-Sample-Problem.. In B. Schölkopf, J. C. Platt & T. Hofmann (eds.), NIPS (p./pp. 513-520), : MIT Press. ISBN: 0-262-19568-2
6. Kang, G., Jiang, L., Yang, Y. & Hauptmann, A. G. (2019). Contrastive Adaptation Network for Unsupervised Domain Adaptation.. CoRR, abs/1901.00976.
7. Ganin, Y. & Lempitsky, V. (2014). Unsupervised Domain Adaptation by Backpropagation
8. Carlucci, F. M., Porzi, L., Caputo, B., Ricci, E. & Bulò, S. R. (2017). AutoDIAL: Automatic Domain Alignment Layers.. ICCV (p./pp. 5077-5085), : IEEE Computer Society. ISBN: 978-1-5386-1032-9
9. Chang, W.-L., Wang, H.-P., Peng, W.-H. & Chiu, W.-C. (2019). All About Structure: Adapting Structural Information Across Domains for Boosting Semantic Segmentation.. CVPR (p./pp. 1900-1909), : Computer Vision Foundation / IEEE.
10. Avants, B. B., Yushkevich, P. A., Pluta, J., Minkoff, D., Korczykowski, M., Detre, J. A. & Gee, J. C. (2010). The optimal template effect in hippocampus studies of diseased populations.. NeuroImage, 49, 2457-2466.
11. Avants, B. B., Tustison, N. J., Song, G., Cook, P. A., Klein, A. & Gee, J. C. (2011). A reproducible evaluation of ANTs similarity metric performance in brain image registration.. NeuroImage, 54, 2033-2044.
12. Fonov, V. S., Evans, A. C., Botteron, K. N., Almli, C. R., McKinstry, R. C. & Collins, D. L. (2011). Unbiased average age-appropriate atlases for pediatric studies.. NeuroImage, 54, 313-327.
13. VS Fonov, AC Evans, RC McKinstry, CR Almli and DL Collins, Unbiased nonlinear average age-appropriate brain templates from birth to adulthood, NeuroImage, Volume 47, Supplement 1, July 2009, Page S102 Organization for Human Brain Mapping 2009 Annual Meeting, DOI: http://dx.doi.org/10.1016/S1053-8119(09)70884-5
14. V. M. Campello et al., "Multi-Centre, Multi-Vendor and Multi-Disease Cardiac Segmentation: The M&Ms Challenge," in IEEE Transactions on Medical Imaging, doi: 10.1109/TMI.2021.3090082.
15. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S. & Schiele, B. (2016). The Cityscapes Dataset for Semantic Urban Scene Understanding. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June.