Is Your Style Transfer Doing Anything Useful? An Investigation Into Hippocampus Segmentation and the Role of Preprocessing

Hoda Kalabizadeh[1], Ludovica Griffanti[2], Pak-Hei Yeung[3], Natalie Voets[4], Grace Gillis[2], Clare Mackay[2], Ana IL Namburete[1], Nicola K Dinsdale[1], and Konstantinos Kamnitsas[5]

[1] Department of Computer Science, University of Oxford, Oxford, UK
[2] Department of Psychiatry, University of Oxford, Oxford, UK
[3] College of Computing and Data Science, Nanyang Technological University, Singapore
[4] Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, UK
[5] Department of Engineering Science, University of Oxford, Oxford, UK
hoda.kalabizadeh@cs.ox.ac.uk

Abstract. Brain atrophy assessment in MRI, particularly of the hippocampus, is commonly used to support diagnosis and monitoring of dementia. Consequently, there is a demand for accurate automated hippocampus quantification. Most existing segmentation methods have been developed and validated on research datasets and, therefore, may not be appropriate for clinical MR images and populations, leading to potential gaps between dementia research and clinical practice. In this study, we investigated the performance of segmentation models trained on research data that were *style-transferred* to resemble clinical scans. Our results highlighted the importance of intensity normalisation methods in MRI segmentation, and their relation to domain shift and style-transfer. We found that whilst normalising intensity based on min and max values, commonly used in generative MR harmonisation methods, may *create* a need for style transfer, Z-score normalisation effectively maintains style consistency, and optimises performance. Moreover, we show for our datasets spatial augmentations are more beneficial than style harmonisation. Thus, emphasising robust normalisation techniques and spatial augmentation significantly improves MRI hippocampus segmentation.

Keywords: Style Transfer · Hippocampus Segmentation · Dementia

1 Introduction

Many neurodegenerative diseases cause volumetric atrophy in the region of the hippocampus [1], including Alzheimer's disease (AD). AD is clinically characterised by a progressive decline in cognitive function with diagnosis and monitoring of the disease commonly including the assessment of hippocampal atrophy

N.K. Dinsdale and K. Kamnitsas — Equal contribution.

in brain MRI scans [2]. Specifically, the volume of the hippocampus, is often measured, either through manual or automated segmentation.

Manual segmentation requires large amounts of time and expert knowledge and suffers from inter-rater variability. Therefore, there is a demand for accurate automated hippocampus segmentation methods [3]. Among different types of techniques, deep learning (DL) based methods show great promise for the segmentation of the hippocampus, outperforming traditional atlas-based approaches [4], [5]. However, training CNNs generally requires the availability of manual labels, limiting the applicability in clinical practice. Furthermore, due to differences in image acquisitions and patient demographics, models trained on research datasets are unlikely to generalise to clinical populations. Therefore, there is a need to overcome this *domain shift* between the source (research) and target (clinical) dataset, enabling the development of segmentation models for clinical scans without requiring segmentation labels.

To address general domain shift, data augmentation is a commonly used technique for artificially enhancing the diversity of training data, to increase model generalisability and robustness. Augmentation has been shown to improve downstream segmentation performance across brain imaging studies [6]. However, augmentation requires the identification and modelling of differences between domains, which is non-trivial in the presence of varying populations and scanner technologies. Another related field of research is image-to-image (I2I) translation, which is an approach that aims to learn the mapping between different visual domains, mostly based on generative models. For instance, Pix2pix [7] utilises a conditional GAN to map between image domains, but relies on pixel-to-pixel correspondence, limiting its applicability to MR images from different sites. CycleGAN [8] overcomes the need for paired data using cycle consistency.

MR harmonisation approaches, e.g., [9] are based on I2I methods, aiming to overcome the style-based domain shifts associated with differing acquisition scanners while maintaining the underlying anatomy.

Therefore, in this study, we aim to investigate which techniques are effective for overcoming the *domain shift* between our source (research) dataset and target (clinical) dataset for the task of MR hippocampus segmentation. Our contributions are as follow:

- We demonstrate the use of a 2-stage pipeline for generating style-transferred images that are subsequently used to train a hippocampus segmentation model.
- We explore the impact on downstream hippocampus segmentation performance of different preprocessing and augmentation approaches.
- We show that the use of appropriate normalisation (*i.e.* Z-score normalisation) and spatial augmentation (*i.e.* paired affine registration) can lead to substantial improvements on downstream hippocampus segmentation performance, even without a sophisticated style transfer pipeline.

The findings of this study may shed light on the importance of developing a robust preprocessing pipeline for MR hippocampus segmentation in future studies.

2 Methods

To overcome the domain shift between the research and clinical data, we implemented a 2-stage approach, formed of a style transfer (ST) network followed by a segmentation network. A schematic of this pipeline is shown in Figure 1. We assumed access to a source (research) dataset, $\mathcal{D}_s = \{\boldsymbol{X}_s, \boldsymbol{Y}_s\}$, and an unlabelled target (clinical) dataset, $\mathcal{D}_t = \{\boldsymbol{X}_t\}$, to first train a style transfer model that generates source images in the style of target images, $\tilde{\boldsymbol{X}}_s = G(\boldsymbol{X}_s, \boldsymbol{X}_t)$, following which we used the ST images to train a segmentation model $f(\tilde{\boldsymbol{X}}_s, \boldsymbol{Y}_s)$, such that the performance for $\mathcal{D}_t$ is maximised.

2.1 Style Transfer: Style-Encoding GAN

We utilised the Style-Encoding GAN (SE-GAN) [9] for our style transfer network. Similarly to StarGANv2 [10], it is formed of a single generator (G), discriminator (D), mapping network (M) and a style encoder (E). During training, SE-GAN trains G to generate diverse images corresponding to a single image slice $\boldsymbol{x} \in \boldsymbol{X}$ using a style code c, provided by either M or E. Consequently, the generator G translates an input image, $\boldsymbol{x}$, into an output image, $\tilde{\boldsymbol{x}} = G(\boldsymbol{x}, c)$, that is reflective of the style of c. To validate the successful injection of c into the output image $\tilde{\boldsymbol{x}}$, E is used to extract the style code from images. The style code is a 1×64 vector, allowing E to produce diverse style codes from different images. Moreover, the discriminator D learns to classify images as real or fake, as produced by $G(\boldsymbol{x}, c)$. In our experiments, G is used to synthesise output images $\tilde{\boldsymbol{x}}_{\boldsymbol{s}}$ based on source images $\boldsymbol{x}_s$ that are reflecting the style c extracted from various reference images in $\boldsymbol{X}_t$. Finally, 3D volumes can be reconstructed by stacking the 2D slices. The network is trained using the loss introduced in [10], formed of an adversarial loss L_{GAN}, cycle consistency loss L_{cyc}, style reconstruction loss L_{sty} and diversification loss L_{div}, weighted by λ_{cyc}, λ_{sty} and λ_{div} respectively, resulting in the following objective function:

$$L(G, M, E, D) = L_{GAN} + \lambda_{cyc} L_{cyc} + \lambda_{sty} L_{sty} - \lambda_{div} L_{div} \tag{1}$$

2.2 Segmentation: U-Net

The second stage of the framework is the training of the segmentation network, f, for which we used a 3D U-Net [11]. The network is trained with a Dice loss, L_{dice}, using the style transformed source data such that:

$$L_{seg}(\tilde{\boldsymbol{X}}_s, \boldsymbol{Y}_s) = L_{dice}(f(\tilde{\boldsymbol{X}}_s), \boldsymbol{Y}_s). \tag{2}$$

2.3 Investigating Preprocessing

Registration: To mitigate content shift, defined as variations in anatomical alignment between brain scans, we investigated the impact of registration on

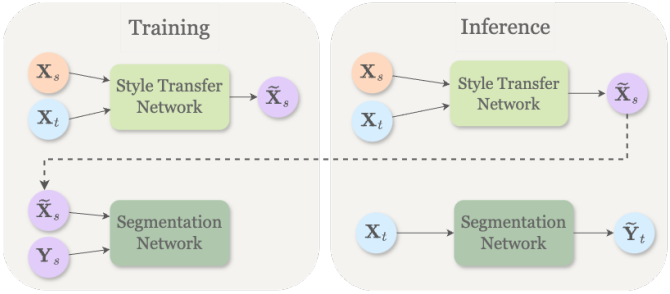


Fig. 1: A schematic of the proposed 2-stage pipeline.

both style transfer and downstream segmentation performance. To this end, we conducted registration, using both 6 (rigid-body) and 12 (affine) degrees of freedom (DOF) and compared two main approaches. (1) **MNI-Reg**: A spatial harmonisation approach, where both source and target images are registered to a standard space, (2) **Paired-Reg**: A spatial augmentation approach, where every source image is registered to every target image.

Normalisation: In most ST models intensity normalisation is performed during training, through linear scaling of the range of intensities from [Min, Max] to a pre-defined range such as [0,1], which we call Min-Max normalisation. This approach, however, is often unsuitable for MRI images as they can have varying intensity distributions from different acquisitions, leading to inconsistent normalisation results. Additionally, intensity outliers common in MRI data can skew normalisation. We investigated the performance of Min-Max normalisation and explored the impact of applying Z-score normalisation on a per-subject basis.

3 Experimental Setup

3.1 Datasets

Research Dataset: The HarP dataset [12] was used as the labelled *research dataset*, consisting of 135 T1-weighted MRI volumes (cognitively normal controls, MCI and AD patients) from a range of scanners, and corresponding hippocampus masks [13]. All MRIs were registered to MNI-space.

Clinical Dataset: We used a dataset from the Oxford Brain Health Clinic (OBHC) [14] as our *clinical dataset*, representing the unlabelled target domain. It includes 29 patients referred to a memory clinic, who agreed to the use of data for research. The lack of strict inclusion criteria typical of a dementia research study, makes this dataset representative of real-world memory clinic patients. The scans were collected using a 3T Siemens scanner. Hippocampi were manually annotated by a clinician. BHC labels were used only for model evaluation, not for training.

Figure 2 compares the hippocampal volumes between the research (HarP) and clinical (BHC) populations. It can be seen that generally the research population have larger hippocampal volumes than the clinical group. This difference

can probably be attributed to dementia research typically recruiting patients that tend to be younger and have less hippocampal atrophy [15].

3.2 Preprocessing

For anonymisation, the BHC scans were brain extracted and thus we performed brain extraction on the HarP dataset. N4 bias field correction was used to correct for low-frequency intensity non-uniformity. Images were split into left and right hemispheres for training. Data registration followed Section 2.3.

3.3 Implementation Details

For training the ST network, 132 HarP images were used as the source images, and 20 randomly selected BHC images were used as the references, using a learning rate of 10^{-4} and the Adam optimiser. For Equation 1, we set $\lambda_{cyc} = 10$, $\lambda_{sty} = 1$ and $\lambda_{div} = 1$, as suggested by [9]. Moreover, 100 HarP images were used for training and then the trained ST network was used to generate a style-transferred image for each HarP image (N=32) in the style of each BHC image (N=20) resulting in 640 style-transferred images, which were used to train the segmentation model.

The chosen U-Net architecture network had four downsampling and upsampling layers, whereby each layer was formed of a convolutional layer, a ReLU activation function and a batch normalisation layer. The depth, defined as the number of convolutions, doubled between each layer, starting with 4. The U-Nets were trained using a learning rate of 10^{-3}, and the Adam optimiser. The training was conducted using 3-fold cross-validation and tested on 9 BHC patients (18 hippocampi) that were not used during the training or validation. Training was conducted using an Nvidia A10 GPU. The ST and segmentation networks required an average training time of 35 hours and 12 hours, respectively. However, once training was complete, the segmentation model's testing, or inference, took only a few seconds per scan, making it suitable for clinical applications.

Table 1: DSC for segmentation methods on HarP (Source) and OBHC (Target). N is the number of test hippocampi (i.e., 2× number of patients). * UDA test sizes were smaller due to training on a sample of unlabelled OBHC.

Method	**HarP (N=64)**	**OBHC (N=58)**
FreeSurfer	0.701 ± 0.049	0.625 ± 0.217
SynthSeg	0.801 ± 0.045	0.732 ± 0.070
FIRST	0.810 ± 0.031	0.758 ± 0.116
Hippodeep	0.829 ± 0.031	0.752 ± 0.062
U-Net	0.854 ± 0.048	0.670 ± 0.171
Basic Aug	0.860 ± 0.041	**0.783 ± 0.051**
MRI Aug	0.854 ± 0.040	0.764 ± 0.055
UDA*	**0.863 ± 0.041**	0.742 ± 0.078

4 Results & Discussion

4.1 Domain Shift

First, to establish the domain shift between the datasets, we tested publicly available out-of-the-box (OOB) tools: FSL FIRST [16], FreeSurfer [17], SynthSeg [5], Hippodeep [18], as well as U-Net based approaches: a U-Net trained solely on HarP, basic augmentation (affine transforms, flips, noise, intensity changes), MRI-specific augmentation (motion, bias field). We also compared with adversarial unsupervised domain adaptation, approach shown potent for tackling domain shift in medical imaging [19], and specifically the model developed in [20] (UDA).

Table 1 shows the dice score (DSC) values for the different approaches. The OOB models all performed better on our research population compared to our clinical population, achieving maximum DSC of 0.829 and 0.758, respectively. In particular, FreeSurfer and FIRST had instances of complete failure for the OBHC data (DSC = 0). For most patients, UDA was comparable to the augmentation methods, however, when examining the worst-case scenarios, UDA led to particularly low dice scores for certain individuals. Data augmentation, thus, proved to be the most effective approach for performance enhancement. These findings demonstrate the limitations of existing methods and highlight the potential value of exploring more sophisticated data augmentation approaches.

4.2 Registration

We then explored the effect of the choice of registration approach. Figure 2 shows the effect of registration on the hippocampal volumes: as rigid registration only involves translation and rotation for brain alignment, there is no change between the original HarP volumes and those registered to OBHC. However, affine registration performs global scaling, resulting in a slight increase in volume of the registered HarP hippocampi. Although the OBHC registration targets have distinctly smaller hippocampal volumes, they have, on average, larger whole-brain volumes, leading to larger hippocampi in the registered HARP images.

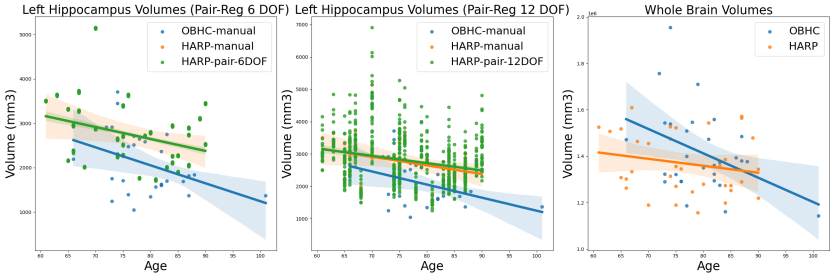


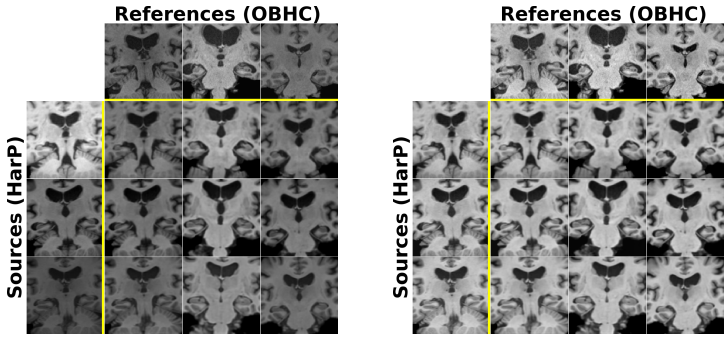
Fig. 2: Volume against age plots for HarP and OBHC: left hippocampus volumes for Paired-Reg 6 DOF (left), 12 DOF (middle), and whole brain volumes (right).

4.3 Style Transfer vs Normalisation

Table 2 shows the segmentation performance for models trained on Z-score and Min-Max normalised data, and tested on the 9 unseen OBHC patients (18 hippocampi), for a range of evaluation metrics, namely the Dice score (DSC), Hausdorff distance (HD), and Relative Absolute Volume Difference (RAVD). As a supervised benchmark, we trained directly on the OBHC dataset (20 labelled patients), which achieved an average DSC of 0.744. By comparison, simply training on the Min-Max normalised HarP dataset achieved a mean DSC of 0.616, clearly indicating presence of a domain shift. The results of training with the U-Net on the HarP dataset with the different registration schemes, normalisation schemes and the use of ST can then be seen. MNI-Reg-6 ST led to a 4% increase in performance, with a DSC of 0.651. Z-score normalisation outperformed Min-Max normalisation across the experiments. Without Z-score normalisation, a noticeable style shift exists, which can be slightly mitigated by training on style-transferred images (MNI-Reg-6 ST). However, implementing Z-score normalisation effectively reduces this style shift, increasing performance to levels similar to a model trained on target data, while the benefits offered by style transfer are reduced. Following this, the impact of mitigating content shift using affine registration was evaluated, specifically employing the paired registration approach (Table 2). A significant increase in segmentation performance is observed through augmenting the data with paired registration (Paired-Reg-12), achieving an average DSC of 0.780 without ST and 0.787 with. Standard augmentations further improved the performance, achieving the highest DSC of 0.797 (Paired-Reg-12 ST + Aug). The source, reference and style-transferred images, generated by ST networks trained on affine paired registered images (Paired-Reg-12 ST) using both normalisation approaches, have been visualised in Figure 3. The figures reveal that Min-Max normalisation tends to highlight the style transfer effect more visibly than Z-score normalisation. This difference arises because Min-Max normalisation is sensitive to extreme values in MRI data, which can distort the results. In contrast, Z-score normalisation is more robust to such outliers. Thus, the differences between

Table 2: Segmentation results using different normalisation and registrations. N is the number of train hippocampi (i.e. 2× number of patients)

Train Data	**Train N**	**Norm**	**Min DSC ↑**	**Avg DSC ↑**	**95 % HD ↓**	**RAVD↓**
OBHC	40	Z-score	0.617	0.744±0.016	3.865±1.357	18.388±1.694
HarP	64	Min-Max	0.480	0.616±0.037	5.264±0.970	97.972±9.456
MNI-Reg-6 ST	1,280	Min-Max	0.521	0.651±0.015	4.036±0.06	83.033±2.037
HarP	64	Z-score	0.674	0.746±0.009	6.013±1.297	26.326±1.856
MNI-Reg-6 ST	1,280	Z-score	0.688	0.757±0.008	5.072±1.391	20.736±1.906
Paired-Reg-12	1,280	Z-score	0.703	0.780±0.005	3.452±0.237	20.009±0.913
Paired-Reg-12 ST	1,280	Z-score	0.672	0.787±0.004	3.169±0.250	23.794±4.033
Paired-Reg-12 ST + Aug	1,280	Z-score	0.719	0.797±0.003	3.133±0.143	27.509±3.724



(a) Min-Max Normalisation (b) Z-score Normalisation

Fig. 3: Generated ST images for a given source (first column) and reference (first row), using a) Min-Max and b) Z-score normalisation.

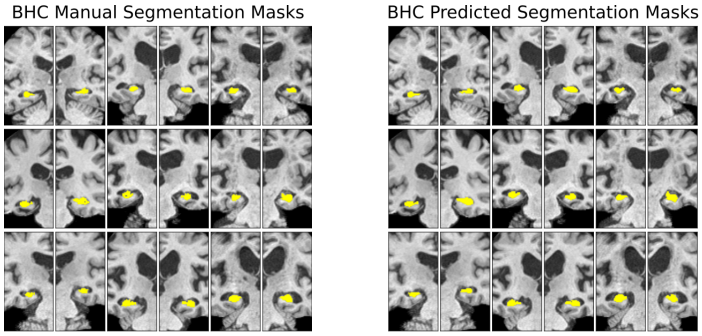


Fig. 4: Manual and predicted segmentation masks for the OBHC test data.

the figures likely reflect variations in intensity ranges rather than style transfer performance. This difference is further demonstrated by the intensity distributions plotted in the Supplementary Material. Figure 4 provides a qualitative comparison between the manual segmentations and the best performing model predictions (Paired-Reg-12 ST with augmentations) on the OBHC test data.

5 Conclusion

In conclusion, we implemented a 2-stage pipeline consisting of a ST and a segmentation network. Our experimental findings underscored the significance of normalisation methods in MRI augmentation and segmentation tasks. While experiments with Min-Max normalisation may suggest a style shift and the potential benefits of style transfer, this interpretation may be misleading and is a result of inappropriate normalisation. Our findings indicate that Z-score normalisation negates the necessity for style transfer by effectively maintaining style consistency in MRI data, thereby optimising segmentation performance directly. More-

over, for the task of hippocampus segmentation, our results demonstrate that mitigating the content shift using a spatial augmentation approach (i.e. Paired-Reg 12 DOF) can be far more beneficial than a spatial harmonisation approach, such as aligning all images to MNI. The improved performance may be attributed to the spatial diversity introduced by the augmentation, which enhances segmentation robustness. Thus, prioritising robust normalisation techniques and appropriate spatial augmentation can lead to substantial improvements in the generalisability of MRI segmentation. Future studies may, thus, benefit from considering spatial augmentation, akin to those currently employed in style transfer, to achieve further improvements in hippocampus segmentation performance.

Acknowledgments. The authors are grateful for support from: the University of Oxford Department of Computer Science Scholarship (HK), the Bill and Melinda Gates Foundation (NKD, AILN) and the Presidential Postdoctoral Fellowship (Nanyang Technological University) (PHY). We are grateful to the operations team of the OBHC. The OHBC data collection and analysis is supported by the NIHR Oxford Health Biomedical Research Centre (NIHR203316) - a partnership between the University of Oxford and Oxford Health NHS Foundation Trust, the NIHR Oxford Cognitive Health Clinical Research Facility, and the Wellcome Centre for Integrative Neuroimaging (203139/Z/16/Z, 203139/A/16/Z). The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care. For the purpose of open access, the authors have applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Disclosure of Interests. The authors have no competing interests to declare.

References

[1] L. Minkova, A. Habich, J. Peter, C. P. Kaller, S. B. Eickhoff, and S. Klöppel, "Gray matter asymmetries in aging and neurodegeneration: A review and meta-analysis," *Human Brain Mapping*, vol. 38, no. 12, pp. 5890–5904, 2017, ISSN: 1097-0193. DOI: 10.1002/hbm.23772.

[2] G. M. McKhann, D. S. Knopman, H. Chertkow, *et al.*, "The diagnosis of dementia due to alzheimer's disease: Recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease," *Alzheimer's & dementia : the journal of the Alzheimer's Association*, vol. 7, no. 3, pp. 263–269, May 2011, ISSN: 1552-5260. DOI: 10.1016/j.jalz.2011.03.005.

[3] E. Balboni, L. Nocetti, C. Carbone, *et al.*, "The impact of transfer learning on 3d deep learning convolutional neural network segmentation of the hippocampus in mild cognitive impairment and alzheimer disease subjects," *Human Brain Mapping*, vol. 43, no. 11, pp. 3427–3438, 2022, ISSN: 1097-0193. DOI: 10.1002/hbm.25858.

[4] N. K. Dinsdale, M. Jenkinson, and A. I. L. Namburete, "Spatial warping network for 3d segmentation of the hippocampus in MR images," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, D. Shen, T. Liu, T. M. Peters, *et al.*, Eds., Cham: Springer International Publishing, 2019, pp. 284–291, ISBN: 978-3-030-32248-9. DOI: 10.1007/978-3-030-32248-9_32.

[5] B. Billot, D. N. Greve, O. Puonti, *et al.*, "SynthSeg: Segmentation of brain MRI scans of any contrast and resolution without retraining," *Medical Image Analysis*, vol. 86, p. 102 789, May 1, 2023, ISSN: 1361-8415. DOI: 10.1016/j.media.2023.102789.

[6] F. Garcea, A. Serra, F. Lamberti, and L. Morra, "Data augmentation for medical imaging: A systematic literature review," *Computers in Biology and Medicine*, vol. 152, p. 106 391, Jan. 1, 2023, ISSN: 0010-4825. DOI: 10.1016/j.compbiomed.2022.106391.

[7] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, *Image-to-image translation with conditional adversarial networks*, version: 1, Nov. 21, 2016. DOI: 10.48550/arXiv.1611.07004. arXiv: 1611.07004[cs].

[8] H. Yang, J. Sun, A. Carass, *et al.*, "Unpaired brain MR-to-CT synthesis using a structure-constrained CycleGAN," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, D. Stoyanov, Z. Taylor, G. Carneiro, *et al.*, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2018, pp. 174–182, ISBN: 978-3-030-00889-5. DOI: 10.1007/978-3-030-00889-5_20.

[9] M. Liu, A. H. Zhu, P. Maiti, *et al.*, "Style transfer generative adversarial networks to harmonize multisite MRI to a single reference image to avoid overcorrection," *Human Brain Mapping*, vol. 44, no. 14, pp. 4875–4892, 2023, ISSN: 1097-0193. DOI: 10.1002/hbm.26422.

[10] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image

translation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, ISSN: 2575-7075, Jun. 2018, pp. 8789–8797. DOI: 10.1109/CVPR.2018.00916.

[11] O. Ronneberger, P. Fischer, and T. Brox, *U-net: Convolutional networks for biomedical image segmentation*, May 18, 2015. DOI: 10.48550/arXiv.1505.04597.

[12] M. Boccardi, M. Bocchetta, L. G. Apostolova, *et al.*, "Delphi definition of the EADC-ADNI harmonized protocol for hippocampal segmentation on magnetic resonance," *Alzheimer's & Dementia*, vol. 11, no. 2, pp. 126–138, 2015, ISSN: 1552-5279. DOI: 10.1016/j.jalz.2014.02.009.

[13] M. Boccardi, M. Bocchetta, F. C. Morency, *et al.*, "Training labels for hippocampal segmentation based on the EADC-ADNI harmonized hippocampal protocol," *Alzheimer's & Dementia*, vol. 11, no. 2, pp. 175–183, 2015, ISSN: 1552-5279. DOI: 10.1016/j.jalz.2014.12.002.

[14] M. C. O'Donoghue, J. Blane, G. Gillis, *et al.*, "Oxford brain health clinic: Protocol and research database," *BMJ Open*, vol. 13, no. 8, e067808, Aug. 1, 2023, Publisher: British Medical Journal Publishing Group Section: Neurology, ISSN: 2044-6055, 2044-6055. DOI: 10.1136/bmjopen-2022-067808.

[15] A. Thorogood, A. Mäki-Petäjä-Leinonen, H. Brodaty, *et al.*, "Consent recommendations for research and international data sharing involving persons with dementia," *Alzheimer's & Dementia*, vol. 14, no. 10, pp. 1334–1343, Oct. 1, 2018, ISSN: 1552-5260. DOI: 10.1016/j.jalz.2018.05.011.

[16] B. Patenaude, S. M. Smith, D. N. Kennedy, and M. Jenkinson, "A bayesian model of shape and appearance for subcortical brain segmentation," *NeuroImage*, vol. 56, no. 3, pp. 907–922, Jun. 1, 2011, ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2011.02.046.

[17] B. Fischl, D. H. Salat, E. Busa, *et al.*, "Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain," *Neuron*, vol. 33, no. 3, pp. 341–355, Jan. 31, 2002, ISSN: 0896-6273. DOI: 10.1016/s0896-6273(02)00569-x.

[18] B. Thyreau, K. Sato, H. Fukuda, and Y. Taki, "Segmentation of the hippocampus by transferring algorithmic knowledge for large cohort processing," *Medical Image Analysis*, vol. 43, pp. 214–228, Jan. 1, 2018, ISSN: 1361-8415. DOI: 10.1016/j.media.2017.11.004.

[19] K. Kamnitsas, C. Baumgartner, C. Ledig, *et al.*, "Unsupervised domain adaptation in brain lesion segmentation with adversarial networks," in *Information Processing in Medical Imaging*, M. Niethammer, M. Styner, S. Aylward, *et al.*, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2017, pp. 597–609, ISBN: 978-3-319-59050-9. DOI: 10.1007/978-3-319-59050-9_47.

[20] N. K. Dinsdale, M. Jenkinson, and A. I. L. Namburete, "Deep learning-based unlearning of dataset bias for MRI harmonisation and confound removal," *NeuroImage*, vol. 228, p. 117 689, Mar. 1, 2021, ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2020.117689.