MDD-UNet: Domain Adaptation for Medical Image Segmentation with Theoretical Guarantees, a Proof of Concept

Asbjørn Munk*[1,2], Ao Ma[3], and Mads Nielsen[1,2]

[1]Pioneer Centre for AI
[2]University of Copenhagen
[3]Southwestern University of Finance and Economics

Abstract

The current state-of-the art techniques for image segmentation are often based on U-Net architectures, a U-shaped encoder-decoder networks with skip connections. Despite the powerful performance, the architecture often does not perform well when used on data which has different characteristics than the data it was trained on. Many techniques for improving performance in the presence of *domain shift* have been developed, however typically only have loose connections to the theory of domain adaption. In this work, we propose an unsupervised domain adaptation framework for U-Nets with theoretical guarantees based on the Margin Disparity Discrepancy [1] called the MDD-UNet. We evaluate the proposed technique on the task of hippocampus segmentation, and find that the MDD-UNet is able to learn features which are domain-invariant with no knowledge about the labels in the target domain. The MDD-UNet improves performance over the standard U-Net on 11 out of 12 combinations of datasets. This work serves as a proof of concept by demonstrating an improvement on the U-Net in it's standard form without modern enhancements, which opens up a new avenue of studying domain adaptation for models with very large hypothesis spaces from both methodological and practical perspectives. Code is available at https://github.com/asbjrnmunk/mdd-unet.

1 Introduction

In medical image analysis data distributions vary considerably across equipment, patient groups, and scanning protocols [2]. Since labeling medical images typically involves labor-intensive participation of specialists, available labeled data is often limited. This is a key challenge in medical image segmentation [3], since models typically fail at generalizing to data which is different from the specific setup of the training data, while manually labeling data from each new test domain is infeasible [4].

One solution to tackles this problem is unsupervised domain adaptation (UDA) [7]. In UDA the goal is to transfer knowledge learned from the source domain to a similar, yet distinct, target domain, while only assuming labels from the source domain.

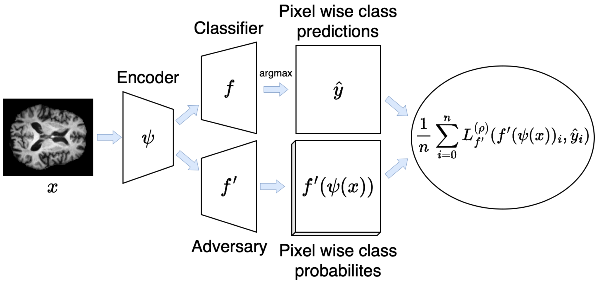
Classifier
Pixel wise class predictions
Encoder
f
argmax
$\hat{y}$
ψ
x
f'
$f'(\psi(x))$
$\frac{1}{n}\sum_{i=0}^{n} L_{f'}^{(\rho)}(f'(\psi(x))_i, \hat{y}_i)$
Adversary
Pixel wise class probabilites

Figure 1. Margin Disparity. The calculation of $\text{disp}_{D,\psi}^{(\rho)}(f', f)$, a measure of disparity between two classifiers, f and f'. This measure importantly works for any classifier, enabling us to apply this directly to medical segmentation. The loss L denotes the *margin loss* up to some maximal margin $\rho > 0$. The final disparity is the average disparity over all pixels in the input.

While important technical advances have been developed in UDA, methodologies are generally not theoretically well-founded. At the same time, remarkable theoretical advances have been achieved in domain adaptation. In particular, the seminal work by Ben-David et al. [8] introduced the $\mathcal{H}\Delta\mathcal{H}$-divergence as a measurement of discrepancy between two distributions. This allowed rigorous learning bounds to be derived and showed, that to obtain good performance on the target domain, there is an intrinsic trade-off between performance on the source domain and the empirical $\mathcal{H}\Delta\mathcal{H}$-divergence.

Practical methods for domain adaptation seek to exploit this trade-off, for instance DANN [6] employ an adversarial architecture, inspired by GAN [9], in which networks play a *minimax game* seeking to learn a representation of the input where the source and target domains are indistinguishable, while performing well on the source domain. However, the theoretical foundation of DANN is limited to binary classifiers, meaning that for problems such as segmentation, the methodology lacks theoretical guarantees, since the hypothesis spaces of the max-player and min-player are distinctly dissimilar.

Zhang et al. [1] remedy this by proposing a new distribution discrepancy measurement called Margin Disparity Discrepancy (MDD), allowing the deriva-

*Corresponding Author: asmu@di.ku.dk

Proceedings of the 5th Northern Lights Deep Learning Conference (NLDL), PMLR 233, 2024.

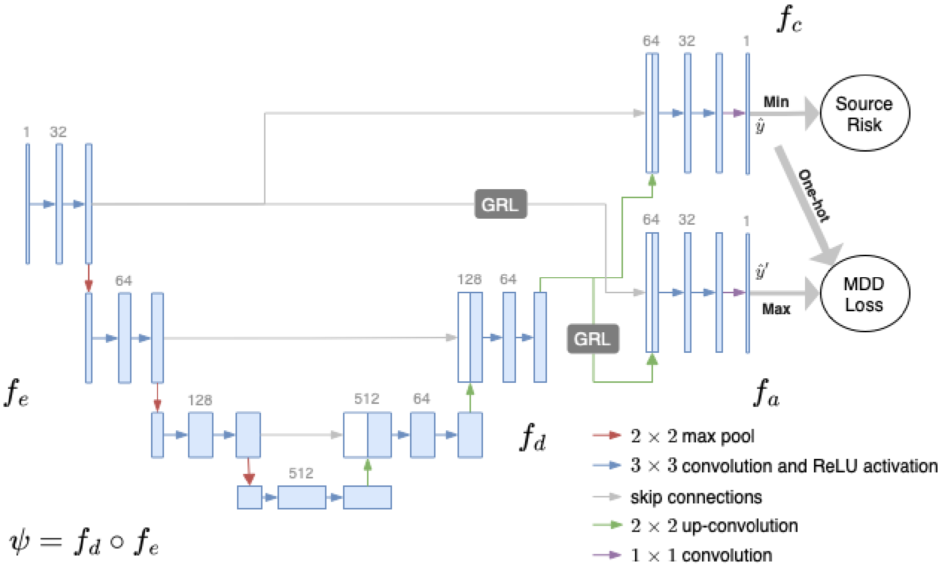

Figure 2. The proposed architecture. The base model is a U-Net [5] combined with the adversarial MDD architecture [1]. Each box denotes an intermediate representation, with the amount of channels denoted by a small number above. GRL is a *Gradient Reversal Layer* [6].

tion of generalization bounds comparable to those of Ben David et al. [8] based on scoring functions and the margin loss. Remarkably, this is seamlessly transformed into a theoretically sound adversarial architecture, with no constrains on the hypothesis spaces of the classifier, achieving considerable improvements over state-of-the-art UDA methods.

While the MDD theory is sound for models with arbitrary hypothesis classes, it is unclear whether it is practical when applied to models with very large hypothesis spaces, such as those used for image segmentation.

Domain Adaptation for biomedical segmentation is currently not well understood theoretically. The theoretical understanding is particularly valuable in the medical domain, since it provides avenues for understanding capacity and limitations of the proposed methodology. This work seeks to investigate whether it is possible to apply MDD to the segmentation problem, by combining the U-Net [5], the architectural foundation for state-of-the-art medical segmentation models, with the MDD and propose a theoretically justified domain adaptation methodology for biomedical image segmentation.

The paper contributes by proposing a new methodology, including a new training procedure with a novel early stopping scheme. The method is evaluated on the task of hippocampus segmentation of brain MRI. We find that the proposed methodology achieve an significant improvement on the base U-Net.

This work is to be considered a proof of concept, and provides an avenue to both further understanding, analyzing and applying Domain Adaptation to the medical domain. The theoretical justification of the proposed methodology, opens up completely new avenues of research, potentially providing fundamental contributions to our understanding of the capabilities and limitations of adversarial Domain Adaptation.

2 Method

This section outlines the proposed method and it's theoretical foundation. We assume input space $\mathcal{X}$ and label space $\mathcal{Y}$, and two distributions S and T over $\mathcal{X} \times \mathcal{Y}$, denoted the *source domain* and *target domain* respectively. The model receives two samples: A labeled sample $\hat{S}$ from S and an unlabeled sample $\hat{T}$ from T_X, the marginal distribution of T over $\mathcal{X}$. The goal is to perform well on the target domain using only $\hat{S}$ and $\hat{T}$.

2.1 Margin Disparity Discrepancy

The theoretical foundation of the proposed method is the *Margin Disparity Discrepancy* (MDD) from Zhang et al. [1]. Let $f : \mathcal{X} \to \mathbb{R}^{|\mathcal{Y}|}$ denote a classifier outputting a *score* for each possible label and $\mathcal{F}$ a family of classifiers. Let $L_f^{(\rho)} : \mathbb{R}^{|\mathcal{Y}|} \times \mathcal{Y} \to \mathbb{R}$ denote the *Margin Loss*, which measures how certain f is at predicting $y \in \mathcal{Y}$ up to some maximal margin $\rho > 0$. This directly induces the *Margin Disparity*, a measure of agreement between two classifiers f and f' measured by the margin loss. Let $h_f : x \mapsto \arg\max_{y \in \mathcal{Y}} f(x)_y$. The Margin Disparity is then for some sample $\hat{D}$ of size m from distribution D over

Table 1. Mean Dice scores with one standard deviation of 3D volumes on the Hippocampus datasets, with each experiment conducted $N = 6$ times.

Source	**Target**	**U-Net**	**MDD-UNet**
Hammers	LPBA40	0.54 ± 0.01	$\mathbf{0.69 \pm 0.02}$
	Oasis	0.65 ± 0.01	$\mathbf{0.71 \pm 0.02}$
	HarP	0.56 ± 0.02	$\mathbf{0.64 \pm 0.01}$
HarP	Hammers	0.67 ± 0.003	$\mathbf{0.68 \pm 0.01}$
	LPBA40	0.49 ± 0.01	$\mathbf{0.54 \pm 0.03}$
	Oasis	0.77 ± 0.01	$\mathbf{0.79 \pm 0.01}$
Oasis	Hammers	0.64 ± 0.02	$\mathbf{0.69 \pm 0.01}$
	LPBA40	0.28 ± 0.1	$\mathbf{0.61 \pm 0.05}$
	HarP	0.6 ± 0.06	$\mathbf{0.72 \pm 0.03}$
LPBA40	Oasis	0.57 ± 0.01	$\mathbf{0.63 \pm 0.01}$
	Hammers	$\mathbf{0.62 \pm 0.01}$	0.60 ± 0.04
	HarP	0.36 ± 0.03	$\mathbf{0.48 \pm 0.09}$

$\mathcal{X}$ given as

$$\text{disp}_{\hat{D}}^{(\rho)}(f', f) := \frac{1}{m} \sum_{i=0}^{m} L_{f'}^{(\rho)}(f'(x_i), h_f(x_i)),$$

a measure of how certain f' is at predicting the same as f averaged over the input image. Figure 1 shows how the Margin Disparity can be calculated from medical images by taking the average over all pixels. The Margin Disparity can be used to formulate a discrepancy metric between two distributions D, D', over $\mathcal{X}$, called the Margin Disparity Discrepancy (MDD). The MDD is defined as

$$\text{mdd}_{f,\mathcal{F}}^{(\rho)}(\hat{S}, \hat{T}) := \sup_{f' \in \mathcal{F}} \left(\text{disp}_{\hat{S}}^{(\rho)}(f', f) - \text{disp}_{\hat{T}}^{(\rho)}(f', f) \right).$$

MDD measures discrepancy between two distributions as the *maximal difference in expected margin loss* of any classifier $f' \in \mathcal{F}$ with respect to f. Importantly, Zhang et al. [1] provides rigorous generalization bounds based on the MDD, showing that there is a trade-off between generalization error and the choice of margin.

A central property of the MDD is that $\mathcal{F}$ can be a family of classifiers which is able to perform medical image segmentation. The MDD is optimizable by following Ganin et al. [6] and introducing a feature transformation, ψ. Applying ψ to the source and target domains, the overall minimization problem can be written as

$$\min_{f,\psi} \text{err}_{\psi(\hat{S})}^{(\rho)}(f) + \text{mdd}_{f,\mathcal{F}}^{(\rho)}(\psi(\hat{S}), \psi(\hat{T})). \qquad (1)$$

This is naturally a *minimax game*, where the goal is to learn a representation, such that the final classification is based on features which are both discriminative and invariant to the change of domains.

2.2 Network Architecture

We combine the MDD with the U-Net. The U-Net is naturally split into *blocks*, each consisting of one or more convolution operations and ReLU activation functions and combined using either max pool in the contracting path and up-convolution in the expanding path. We only consider models applied on 2D data, which can be obtained from 3D volumes by considering each slice independently. We apply MDD to the U-Net by splitting it into four parts:

1. f_c: The classifier. The top block of the expanding path, consisting of two convolution layers and ReLU activations and the final segmentation layer.

2. f_a: The adversary. An exact architectural copy of the classifier.

3. f_d: The decoder. All blocks on the expanding path except for the classifier and adversary.

4. f_e: The encoder. All blocks on the contracting path including the last block until the first up-convolution.

We let $\psi = f_e \circ f_d$ and optimize Equation 1 using an adversarial architecture. ψ is trained through a Gradient Reversal Layer (GRL) (subsection 2.3). The architecture is given in Figure 2.

2.3 Gradient Reversal Layer

The Gradient Reversal Layer (GRL) follows Ganin et al. [6]. In the forward pass, the layer is simply the identity function. In the backwards pass the gradient is multiplied with a negative constant, η, effectively forcing ψ to transform the input into domain invariant features by maximizing the MDD which minimise loss with respect to the parameters passed through the GRL.

2.4 Loss

Since the margin loss is prone to vanishing gradients, we follow [1] and use the cross entropy loss to optimize Equation 1. Let $\sigma : \mathbb{R}^K \to (0,1)^K$ denote the softmax. Let $N = 256^2 \cdot B$ where B is the batch size. For source data $(x^s, y^s) \in \hat{S}$ the classifier, f_c, simply seeks to approximate $\operatorname{err}^{(\rho)}_{\psi(S)}(f)$ using the standard cross entropy loss

$$L^c(x^s, y^s) := -\frac{1}{N}\sum_{i=1}^{N} \log\left[\sigma(f_c(\psi(x^s)))_{y^s}\right]_i. \quad (2)$$

For target data $x^t \in \hat{T}$ the adversary seeks to approximate $\operatorname{mdd}^{(\rho)}_{f,\mathcal{F}}(\psi(\hat{S}), \psi(\hat{T}))$ using

$$L^{a'}(x^s) := -\frac{1}{N}\sum_{i=1}^{N} \log\left[\sigma(f_a(\psi(x^s)))_{h_{f_c}(x^s)}\right]_i \quad (3)$$

$$L^{a''}(x^t) := \frac{1}{N}\sum_{i=1}^{N} \log\left[1-\sigma(f_c(\psi(x^t)))_{h_{f_c}(x^t)}\right]_i \quad (4)$$

where the modification of the cross entropy loss in Equation 4 was introduced in [9] to mitigate the adversarial burden of exploding or vanishing gradients. The total loss of f_a is

$$L^a(x^s, x^t) := -L^{a''}_{\psi,f_a,f_c}(x^t) + \gamma\, L^{a'}_{\psi,f_a,f_c}(x^s). \quad (5)$$

which is combined using a *margin factor*, $\gamma = \exp \rho$. Note that the adversary is completely unsupervised, and instead depends on the classifier f_c. The margin factor γ is treated as a hyperparameter, and is preferred relatively larger, however might lead to exploding gradients for large values. Note that Equation 5 is formulated as a minimization problem, and thus the total objective of the MDD becomes

$$\min_{\psi,f_a,f_c} L^a_{\gamma,\psi,f_a,f_c}(x^s, x^t) + L^c_{f_c,\psi}(x^s, y^s), \quad (6)$$

which can be directly optimized using stochastic gradient descent.

2.5 Pre-training and early stopping

The MDD-UNet is trained by first training a standard U-Net on the source dataset. That is the model $f_c \circ f_d \circ f_e$, trained with the loss given in Equation 2. MDD is then applied by copying the weights of f_c into f_a and training with the loss from the previous section, effectively treating the domain adaptation as a fine-tuning step.

Table 2. Frozen layers. Dice score on the target distribution by epoch on the validation split for different choices of freezing blocks. We define a block as the convolution layers with the same feature map size delimited by max-pool or up-convolution. We count the blocks from left to right, that is the order of the forward pass.

Frozen layers \ Epoch	**2**	**6**	**12**
First encoder block	**0.6**	**0.63**	0.68
First 2 encoder blocks	**0.6**	0.62	**0.69**
First 3 encoder blocks	0.6	0.61	0.63
All of encoder	0.58	0.59	0.62
Last 2 blocks of encoder	0.48	0.28	0.38
Last block of encoder + first block of decoder	0.45	0.36	0.12
First 2 blocks in decoder	0.36	0.01	0.0

Since MDD is applied to a model which has already learned to segment the source dataset Equation 3 is expected to be numerically quite close to zero. Interestingly, this metric is seemingly associated with performance degradation, and tends to increase rapidly right before performance degenerates. Thus, we introduce a new early stopping metric, which importantly can be performed without knowledge about the labels of T: Let $\xi > 0$, we then stop training immediately when

$$L^{a'}_{\psi,f_a,f_c}(x^s) > \xi,$$

for some batch of source data x^s. In practice we used $\xi = 0.02$.

3 Experimental setup

We validate the effectiveness of the proposed methodology on the task of hippocampus segmentation.

3.1 Data

The core data used in this study are T1-weighted MRI volumes from [10]. Labels highlighted the hippocampus, separated into three class labels: left hippocampus, right hippocampus and background. The data consists of four datasets, which are used to represent distributional shift, by choosing different datasets as the source and target domains respectively. The datasets are:

1. **HarP**: 135 MRI scans from the ADNI study [11] of normal, cognitively impaired, and demented subjects (65 female and 70 males) aged 60 to 90. Data was acquired using scanners from GE, Philips, and Siemens, with strengths of 1.5T or 3T.

2. **Hammers**: 30 MRI scans from [12] of healthy subjects (15 female and 15 male) aged 20 to 54. Data was acquired using a 1.5T GE scanner.

3. **Oasis**: 35 MRI scans from the MICCAI 2012 Multi-Atlas Labeling challenge [13, 14] of healthy subjects (22 females and 13 males) aged between 18 and 90. Data was acquired using a 1.5T GE scanner.

4. **LPBA40**: 40 MRI scans from [15] of healthy subjects (22 females and 13 males) aged between 19 and 40. Data was acquired using a 1.5T GE scanner.

3.2 Preprocessing

All volumes were skull stripped using the robust learning-based brain extraction system ROBEX [16], bias field corrected and transformed to the RAS+ orientation. Moreover, the intensities of each volume were limited to the 99th percentile, standardized to have a zero mean and unit variance, and then scaled to the range -1 to 1. Since the network only handles 2D input, volumes were sliced on the coronal dimension and padded to size 256×256.

3.3 Model configurations

The MDD-UNet is compared to the U-Net. U-Nets are first trained for 60 epochs before MDD is applied. Models were trained using Adam [17] with different learning rates for different parts of the MDD-UNets. When applying MDD, we freeze the first two layers of f_e, and apply early stopping with $\xi = 0.02$. When training MDD on top of U-Net trained with augmentations, the MDD did not use augmentations. U-Nets are trained with an initial learning rate of 10^{-3} which is subsequently halved every 10 epochs. When applying MDD the learning rate of f_a was 10^{-6}, the learning rate of f_c was 10^{-3}, f_e and f_d was 10^{-3} multiplied by $\frac{2}{3}$ and $\frac{4}{9}$ respectively. The MDD-UNet used a margin factor of $\gamma = 0.08$, and a GRL constant of $\alpha = 1.4$.

4 Results

The results of our experiment are given in Table 1. The performance of MDD-UNet is a substantial improvement on the base U-Net, obtaining the best performance on 11 out of 12 combinations.

5 Discussion & Limitations

Frozen layers. To analyze the impact of freezing layers, we conducted an experiment looking at the performance with different blocks frozen. We define a block as the convolution layers with the same feature map size delimited by max-pool or up-convolution. We count the blocks from left to right, that is the order of the forward pass. Table 2 denotes the dice score on the target distribution by epoch in the training progress on the validation split. The performance of the U-Net before adding MDD was 0.54 Dice on the target set. Freezing the first two blocks of encoder outperforms all other configurations, in particular any configuration where blocks in the decoder is frozen.

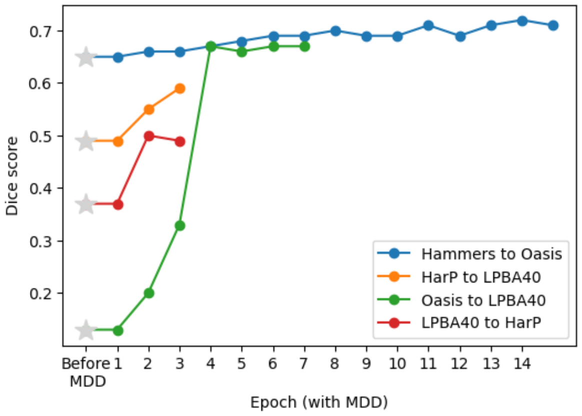

Figure 3. MDD Learning Curves. Learning curves on the target domain for one run on different combination of source and target domains. Dice score is shown after each epoch of applying MDD and the length of each line denote when the training was stopped due to the early stopping mechanism, discussed in subsection 2.5. The learning curves are from one of the 6 runs performed in each combination of datasets. The plot clearly shows that the MDD is effective at mitigating performance loss due to domain shift, as the performance efficiently improves when the MDD is applied.

The frozen layers of the MDD-UNet indicate that the low-level features of the model are more domain invariant in the U-Net than the high level features. Further, since the hypothesis space of the max-player is incredibly large, it can be difficult to find the desired equilibrium between the adversaries. These results showcase how a combination of frozen layers in the beginning of ψ and pre-training, achieves a stable training which allows the MDD to be applied with the early-stopping mechanism.

Effectiveness of the MDD. When the MDD is applied, the performance of the network on the target domain is efficiently improved. Figure 3 shows the learning curves as measured by Dice performance on the target domain by number of epochs of MDD application. When the MDD is applied, the target performance improves dramatically in only a few epochs. The early stopping mechanism reliably stops training exactly when the target performance is best or close to.

Limitations. This work does not claim to establish the MDD-UNet as a state-of-the art domain adaptation methodology, and future work should investigate the interplay with augmentations and other methodological improvements known to improve performance in the presence of domain shift [18]. Further, in this work we focused on demonstrating the effectiveness of the MDD on models working on 2D data. It is left for future work to investigate how the methodology behaves on 3D data, which is common in the medical domain, and modern adaptations of the U-Net [19, 20].

6 Conclusion

In this paper we proposed a domain adaptation methodology with theoretical guarantees based on the U-Net and the MDD. We show that the MDD-UNet outperforms the regular U-Net for segmenting hippocampus data. This work opens the door for further studying the applications of the proposed methodology and importantly the MDD discrepancy metric to the biomedical domain. Further, this work opens the door for analysing biomedical Domain Adaptation theoretically, a completely new avenue of research in the biomedical field.

Acknowledgments

This work was supported by Danish Data Science Academy, which is funded by the Novo Nordisk Foundation (NNF21SA0069429) and VILLUM FONDEN (40516), and by the Pioneer Centre for AI (DNRF grant number P1).

References

[1] Y. Zhang, T. Liu, M. Long, and M. Jordan. "Bridging Theory and Algorithm for Domain Adaptation". In: *International Conference on Machine Learning*. 2019, pp. 7404–7413.

[2] P. Saat, N. Nogovitsyn, M. Y. Hassan, M. A. Ganaie, R. Souza, and H. Hemmati. "A domain adaptation benchmark for T1-weighted brain magnetic resonance image segmentation". In: *Frontiers in Neuroinformatics* 16 (2022). ISSN: 1662-5196. DOI: 10.3389/fninf.2022.919779.

[3] M. Orbes-Arteaga, T. Varsavsky, C. H. Sudre, Z. Eaton-Rosen, L. J. Haddow, L. Sørensen, M. Nielsen, A. Pai, S. Ourselin, M. Modat, P. Nachev, and M. J. Cardoso. "Multi-domain Adaptation in Brain MRI Through Paired Consistency and Adversarial Learning". In: *Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data*. Ed. by Q. Wang, F. Milletari, H. V. Nguyen, S. Albarqouni, M. J. Cardoso, N. Rieke, Z. Xu, K. Kamnitsas, V. Patel, B. Roysam, S. Jiang, K. Zhou, K. Luu, and N. Le. Cham: Springer International Publishing, 2019, pp. 54–62. ISBN: 978-3-030-33391-1.

[4] K. Kamnitsas, C. Baumgartner, C. Ledig, V. Newcombe, J. Simpson, A. Kane, D. Menon, A. Nori, A. Criminisi, D. Rueckert, and B. Glocker. "Unsupervised Domain Adaptation in Brain Lesion Segmentation with Adversarial Networks". In: *Information Processing in Medical Imaging*. Ed. by M. Niethammer, M. Styner, S. Aylward, H. Zhu, I. Oguz, P.-T. Yap, and D. Shen. Cham: Springer International Publishing, 2017, pp. 597–609. ISBN: 978-3-319-59050-9.

[5] O. Ronneberger, P. Fischer, and T. Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Ed. by N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi. Cham: Springer International Publishing, 2015, pp. 234–241. ISBN: 978-3-319-24574-4.

[6] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. March, and V. Lempitsky. "Domain-Adversarial Training of Neural Networks". In: *Journal of Machine Learning Research* 17.59 (2016), pp. 1–35. URL: http://jmlr.org/papers/v17/15-239.html.

[7] H. Guan and M. Liu. "Domain Adaptation for Medical Image Analysis: A Survey". In: *IEEE Transactions on Biomedical Engineering* PP (Oct. 2021), pp. 1–1. DOI: 10.1109/TBME.2021.3117407.

[8] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. "A theory of learning from different domains". In: *Machine Learning* 79.1 (2010), pp. 151–175. DOI: 10.1007/s10994-009-5152-4. URL: https://doi.org/10.1007/s10994-009-5152-4.

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. "Generative Adversarial Nets". In: *Advances in Neural Information Processing Systems*. Ed. by Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger. Vol. 27. Curran Associates, Inc., 2014.

[10] M. M. Ghazi and M. Nielsen. "FAST-AID Brain: Fast and Accurate Segmentation Tool using Artificial Intelligence Developed for Brain". In: (2022). arXiv: 2208.14360 [eess.IV].

[11] C. Jack, M. Bernstein, N. Fox, P. Thompson, G. Alexander, D. Harvey, B. Borowski, P. Britson, J. Whitwell, C. Ward, A. Dale, J. Felmlee, J. Gunter, D. Hill, R. Killiany, N. Schuff, S. Fox-Bosetti, C. Lin, C. Studholme, and M. Weiner. "The Alzheimer's Disease neuroimaging initiative (ADNI): MRI methods". In: *Journal of magnetic resonance imaging : JMRI* 27 (May 2008), pp. 685–91. DOI: 10.1002/jmri.21049.

[12] I. Faillenot, R. A. Heckemann, M. Frot, and A. Hammers. "Macroanatomy and 3D probabilistic atlas of the human insula". In: *NeuroImage* 150 (2017), pp. 88–98. ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2017.01.073. URL: https://www.sciencedirect.com/science/article/pii/S1053811917300964.

[13] D. Marcus, T. Wang, J. Parker, J. Csernansky, J. Morris, and R. Buckner. "Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI Data in Young, Middle Aged, Nondemented, and Demented Older Adults". In: *Journal of cognitive neuroscience* 19 (Oct. 2007), pp. 1498–507. DOI: 10.1162/jocn.2007.19.9.1498.

[14] S. W. B. Landman. "MICCAI 2012: Grand Chal- lenge and Workshop on Multi-atlas Labeling". In: *MICCAI Grand Challenge and Workshop on Multi-Atlas Labeling*. 2012.

[15] D. W. Shattuck, M. Mirza, V. Adisetiyo, C. Hojatkashani, G. Salamon, K. L. Narr, R. A. Poldrack, R. M. Bilder, and A. W. Toga. *Construction of a 3D probabilistic atlas of human cortical structures*. 2008. DOI: https://doi.org/10.1016/j.neuroimage.2007.09.031. URL: https://www.sciencedirect.com/science/article/pii/S1053811907008099.

[16] J. Iglesias, C.-Y. Liu, P. Thompson, and Z. Tu. "Robust Brain Extraction Across Datasets and Comparison With Publicly Available Methods". In: *IEEE transactions on medical imaging* 30 (Sept. 2011), pp. 1617–34. DOI: 10.1109/TMI.2011.2138152.

[17] D. Kingma and J. Ba. "Adam: A Method for Stochastic Optimization". In: *International Conference on Learning Representations* (Dec. 2014).

[18] C. Chen, Q. Dou, H. Chen, J. Qin, and P.-A. Heng. "Synergistic Image and Feature Adaptation: Towards Cross-Modality Domain Adaptation for Medical Image Segmentation". In: *Proceedings of The Thirty-Third Conference on Artificial Intelligence (AAAI)*. 2019, pp. 865–872.

[19] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation". In: *Nature Methods* 18.2 (Feb. 2021), pp. 203–211. ISSN: 1548-7105. DOI: 10.1038/s41592-020-01008-z. URL: https://doi.org/10.1038/s41592-020-01008-z.

[20] H. H. Lee, S. Bao, Y. Huo, and B. A. Landman. "3D UX-Net: A Large Kernel Volumetric ConvNet Modernizing Hierarchical Transformer for Medical Image Segmentation". In: *The Eleventh International Conference on Learning Representations*. 2023. URL: https://openreview.net/forum?id=wsZsjOSytRA.