TUM School of Computation, Information and Technology
Technische Universität München

Efficient and Cross-Domain Deep Learning for Advanced Neuroimage Analysis

Hongwei Li

Vollständiger Abdruck der von der TUM School of Computation, Information and Technology der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz: Prof. Dr. Daniel Rückert

Prüfer*innen der Dissertation:

1. Prof. Dr. Bjoern Menze
2. Prof. Dr. Koen Van Leemput
3. Prof. Tammy Riklin Raviv, Ph.D.

Die Dissertation wurde am 21.12.2022 bei der Technischen Universität München eingereicht und durch die TUM School of Computation, Information and Technology am 19.07.2023 angenommen.

In loving memory of my beloved grandfather and uncle.
怀念已逝去的梁氏李氏亲人。

Abstract

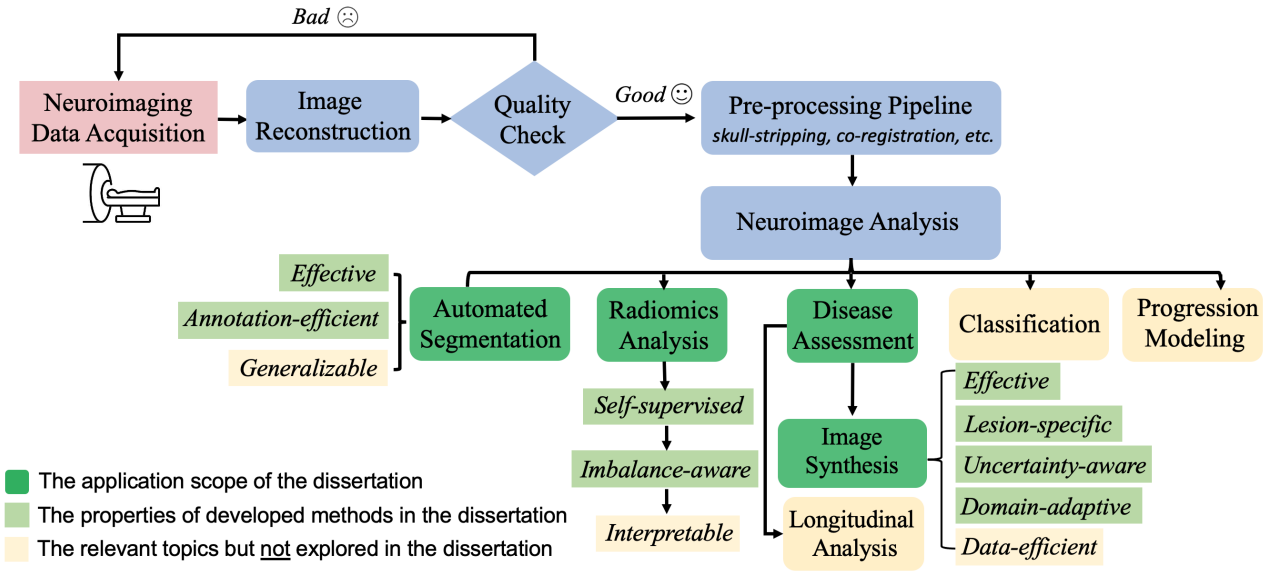
Increasingly large-scale neuroimaging datasets facilitate data-driven domain knowledge modeling. Such modeling can potentially enhance the diagnosis of neurological diseases by providing non-invasive and reliable biomarkers from advanced imaging tools. Deep learning has shown promising progress in various neuroimage analysis tasks. However, its deployment in clinics and laboratories remains challenging due to a combination of labeling scarcity, missing data, and domain shifts between image datasets, forming a high adoption barrier in real-world settings. This dissertation addresses these bottlenecks by developing novel deep-learning methods that effectively and efficiently process neuroimages, acknowledge the imperfection of data, and enable the models to adapt to unseen domains. The contributions of this dissertation are categorized into three neuroimage analysis tasks: (a) image segmentation, (b) image synthesis, and (c) radiomics analysis.

For image segmentation, we develop *effective* and *efficient* approaches for brain structure segmentation and validate them in multi-center settings. Firstly, to segment white matter hyperintensities in FLAIR and T1 images, we introduce an ensemble of multiple independent networks to improve lesion detection. This algorithm was independently evaluated on unseen multi-center datasets and won the 1^{st} place in a grand challenge at MICCAI 2017. Secondly, we propose an ensemble of networks learned from multiple views for the segmentation of the claustrum, a deep gray matter structure. We validate it in a multi-center dataset and observe that it achieves excellent segmentation accuracy compared to human annotators. Thirdly, we explore *deep transfer learning* to significantly reduce annotation effort in a new image domain. We transfer the knowledge learned from one modality to another for the same segmentation task.

For image synthesis, firstly, we develop a *unified* generative adversarial networks (GANs) method for cross-modality synthesis and validate it in a clinical setting, where we observe that the synthetic images improve the diagnosis of multiple sclerosis in brains. It demonstrates that we can augment image information from almost-invisible to visible. Secondly, we enhance the developed GANs approach with two new properties: *lesion-specific* and *uncertainty-aware*. We propose an improved loss function to focus on the lesion region and validate the new loss in a multi-center setting. Among all centers, we consistently demonstrate that synthetic images can improve diagnosis compared to traditional FLAIR images. Importantly, we show that uncertainty maps can detect false positives in synthetic images, thus improving the trustworthiness of the established method. Third, we explore an efficient 3D *unsupervised domain adaptation* method for image synthesis to address the domain shift between source and target domains.

For radiomics analysis, we introduce *contrastive learning* for learning 3D data-driven radiomics in a self-supervised fashion and address the *imbalance issue* in the self-supervised setting. More importantly, we demonstrate that the new radiomics features complement existing ones.

The recent progress of deep learning for neuroimage analysis is still largely confined to highly controlled settings. The work in this dissertation shows that: (1) leveraging data efficiently under limited annotation or without annotations can benefit existing analysis workflows, (2) augmenting image information can improve clinical outcomes, and (3) handling domain shifts might be the key to transforming research achievements in image computing into clinical practice.



A graphical representation of this dissertation's contributions to the field of neuroimaging.

Zusammenfassung

Die Sammlung von immer umfangreicheren Neuroimaging-Daten bietet die Möglichkeit zur datengesteuerten Modellierung von Fachwissen. Eine solche Modellierung kann möglicherweise die Diagnose neurologischer Erkrankungen verbessern, indem sie nicht-invasive und zuverlässige Biomarker aus MR-Bildgebungsinstrumenten liefert. Auf Deep Learning basierende Methoden haben vielversprechende Fortschritte bei der Analyse von Neurobilddaten gezeigt. Ihr Einsatz in Kliniken und Labors bleibt jedoch eine Herausforderung. Eine Kombination aus der Knappheit an Beschriftungen, fehlenden Daten und der Verschiebung zwischen verschiedenen Bilddatensätzen stellt eine hohe Hürde für den Einsatz in der Praxis dar. Diese Dissertation zielt darauf ab, diese Engpässe zu beseitigen, indem neuartige Deep-Learning-Methoden entwickelt werden, die Neurobilddaten effektiv und effizient verarbeiten, die Unvollkommenheit von Datensätzen berücksichtigen und die Modelle dazu bringen, sich an unbekannte Domänen anzupassen. Die Beiträge dieser Dissertation gliedern sich in drei Aufgaben der Neurobildanalyse: (a) Bildsegmentierung, (b) Bildsynthese und (c) Radiomics-Analyse.

Für die Bildsegmentierung entwickeln wir *effektive* und *effiziente* Ansätze für die Segmentierung von Gehirnstrukturen und validieren sie in multizentrischen Einstellungen. Erstens führen wir für die Segmentierung von Hyperintensitäten der weißen Substanz in FLAIR- und T1-Bildern ein Ensemble aus mehreren unabhängigen Netzwerken ein, um die Erkennung von Läsionen zu verbessern. Der Algorithmus wurde unabhängig auf ungesehenen multizentrischen Datensätzen evaluiert und gewann den 1^{st} Platz in einer großen Herausforderung bei MICCAI 2017. Zweitens schlagen wir ein Ensemble von Netzwerken gelernt aus mehreren Ansichten für die Segmentierung von Claustrum (eine tiefe graue Materie Struktur). Wir validieren es in einem multizentrischen Datensatz und stellen fest, dass es im Vergleich zu menschlichen Annotatoren eine ausgezeichnete Segmentierungsgenauigkeit erreicht. Drittens erforschen wir *deep transfer learning*, um den Annotationsaufwand in einem neuen Bildbereich erheblich zu reduzieren. Wir übertragen das von einer Modalität gelernte Wissen auf eine andere für dieselbe Segmentierungsaufgabe.

Für die Bildsynthese entwickeln wir zunächst eine *unified* generative adversarial networks (GANs) Methode zur Durchführung einer modalitätsübergreifenden Synthese und validieren sie in einem klinischen Umfeld, in dem wir beobachten, dass die synthetischen Bilder die Diagnose von Multipler Sklerose in Gehirnen verbessern. Es zeigt sich, dass wir die Bildinformationen von fast unsichtbar zu sichtbar erweitern. Zweitens erweitern wir den entwickelten GAN-Ansatz um zwei neue Eigenschaften: *lesionsspezifisch* und *uncertainty-aware*. Wir schlagen eine verbesserte Verlustfunktion vor, die sich auf die Läsionsregion konzentriert, und validieren den neuen Verlust in einer multizentrischen Umgebung. In allen Zentren konnten wir durchweg zeigen, dass synthetische Bilder die Diagnose im Vergleich zu herkömmlichen FLAIR-Bildern verbessern können. Wichtig ist, dass wir zeigen, dass Unsicherheitskarten falsch-positive Läsionen in synthetischen Bildern erkennen können und somit die Vertrauenswürdigkeit der etablierten Methode verbessern. Drittens erforschen wir eine effiziente 3D *unsupervised domain adaptation*-Methode für die Bildsynthese, um die Domänenverschiebung zwischen Quell- und Zieldomänen anzugehen.

Für die Radiomics-Analyse führen wir das kontrastive Lernen ein, um datengesteuerte 3D-Radiomics auf selbstüberwachte Weise zu erlernen und das Problem des Ungleichgewichts in der selbstüberwachten Umgebung zu lösen. Noch wichtiger ist, dass wir zeigen, dass die neuen Radiomics-Merkmale die bestehenden ergänzen.

Die jüngsten Fortschritte beim Deep Learning für die Neurobildanalyse beschränken sich noch weitgehend auf stark kontrollierte Umgebungen. Die in dieser Arbeit vorgestellte Arbeit zeigt, dass

(1) die effiziente Nutzung von Daten (mit begrenzten Annotationen oder ohne Annotationen) und die Anreicherung von Bildinformationen die klinischen Ergebnisse verbessern können, und (2) der Umgang mit Domänenverschiebungen der Schlüssel zur Übertragung von Forschungsergebnissen in der Bildverarbeitung in die reale klinische Praxis sein könnte.

Acknowledgement

First of all, I would like to express my gratitude to my supervisor, Prof. Bjoern Menze, for his unconditional support, continued encouragement, and enthusiastic supervision during my years in Munich and Zurich. Bjoern has been instrumental in facilitating in-depth discussions and providing me with substantial space to explore, learn, and grow as an independent researcher. He is very 'generative', providing a wide range of directions for new topics and problems while I am more towards a discriminative' style, preferring specific solutions during most discussions. He has been hugely helpful in brainstorming, polishing research ideas, and bringing resources (equally to everyone). I am also thankful for creating such a relaxed, friendly, international, and collaborative culture for his group at TUM. Bjoern was traveling all the time, between Freiburg and Munich before COVID-19, and later between Freiburg and Zurich. Inspired by him, I learned to schedule one-quarter of the research work into traveling.

I would like to thank Prof. Jianguo Zhang and Prof. Wei-Shi Zheng for bringing me to computer vision, especially Jianguo who is one of my closest collaborators and a great advisor in many aspects, initially in Guangzhou, Dundee, and lately in Shenzhen. I am always impressed by his thorough understanding of the micro-concepts of machine learning. I would like to thank Prof. Kuangyu Shi for his continued collaboration and supervision in Munich/Bern. I am especially impressed by his effort to build links between China and Europe in nuclear medicine.

During my Ph.D. program, I have been fortunate to learn from many mentors with diverse clinical research backgrounds. I would like to thank Bene Wiestler for his enthusiastic supervision of many neuroimaging projects (even on weekends). I see his absolute passion for clinical and machine learning research during the interactions. In parallel, I would like to thank Jan S. Kirschke, another extremely hard-working guy who provides radiological insights and contributes to many publications together. I would like to thank Dennis Hedderich and Christian Sorg for showing a piece of the neuroscience field in which imaging tools have broader applications and impacts. I am extremely thankful for the collaborative research culture in Klinikum rechts der Isar and the support from all clinical collaborators, who made this truly interdisciplinary work possible.

Special thanks to my colleagues and friends (*IBBMers*) who created fun in my working space and made this thesis possible, from **A**mir, Anjany, Cadgas, Carolin, Chinmay, Dhritiman, Diana, Fernando, Florian, Giles, Ivan, Izabela, Jana, Johannes, John, Judith, Lina, Lucas, Marie, Markus, Oliver, Rami, Sandeep, Sebastian, Supro, Tamaz, Timo, Xiaobin, Yu to **Y**usuf. Thanks for all the discussions, ideas, and gossip at IMETUM, TranslaTUM, and the BBQ/hotpot sessions.

I would like to thank Pedro, Miguel, and Maria for my internship at Orbem during the pandemic. It was the best and most unforgettable experience to work in such an amazing startup and with wonderful colleagues. It was not the best way to say goodbye due to personal reasons. I sincerely wish you the best of luck in the journey of pushing the boundary of imaging.

I would like to thank Daniel to host me at TUM when I decided to move back from Zurich. It was a new experience to learn how things were organized and developed in a new research chair. Since then, I know many new friends and colleagues and exchanged many research ideas.

Special thanks to my volleyball friends, from **A**lessia, Andy, Antonio, Carmen, Carlos, Dalong, Laura, Mei-Ling, Jason, Lu, to **S**iyuan, for the titles in the tournaments in Scotland and Germany, and for the play and mentally-connected time together. I am not sure if I can play anymore due to my broken knees. But I believe my passion will go on for future years in the rest of my life.

I am indebted to my family members, in particular my parents, my sisters, and my brother for their enduring love and endless support.

Publication list

This cumulative dissertation is based on the following eight *first-author* publications which are grouped by three applications in neuroimage analysis. It covers machine learning topics, including ensemble models, transfer learning, domain adaptation, and self-supervised representation learning.

Work on image segmentation

- **H. Li**, G. Jiang, J. Zhang, R. Wang, Z. Wang, W.-S. Zheng, and B. Menze. "Fully convolutional network ensembles for white matter hyperintensities segmentation in MR images". In: ***NeuroImage*** 183 (2018), pp. 650–665.
- **H. Li**, A. Menegaux, B. Schmitz-Koep, A. Neubauer, F. J. Bäuerlein, S. Shit, C. Sorg, B. Menze, and D. Hedderich. "Automated claustrum segmentation in human brain MRI using deep learning". In: ***Human Brain Mapping*** 42.18 (2021), pp. 5862–5872.
- A. Neubauer*, **H. Li***, J. Wendt, B. Schmitz-Koep, A. Menegaux, D. Schinz, B. Menze, C. Zimmer, C. Sorg, and D. M. Hedderich. "Efficient claustrum segmentation in T2-weighted neonatal brain MRI using transfer learning from adult scans". In: ***Clinical Neuroradiology*** (2022), pp. 1–12.

Work on image synthesis

- **H. Li***, J. C. Paetzold*, A. Sekuboyina, F. Kofler, J. Zhang, J. S. Kirschke, B. Wiestler, and B. Menze. "DiamondGAN: unified multi-modal generative adversarial networks for MRI sequences synthesis". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (**MICCAI**)*. Springer. 2019, pp. 795–803.
- Q. Hu*, **H. Li***, and J. Zhang. "Domain-adaptive 3D medical image synthesis: an efficient unsupervised approach". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (**MICCAI**)* (2022), pp. 495–504.
- T. Finck*, **H. Li***, L. Grundl, P. Eichinger, M. Bussas, M. Mühlau, B. Menze, and B. Wiestler. "Deep-learning generated synthetic double inversion recovery images improve multiple sclerosis lesion detection". In: ***Investigative Radiology*** 55.5 (2020), pp. 318–323.
- T. Finck*, **H. Li***, S. Schlaeger, L. Grundl, N. Sollmann, B. Bender, E. Bürkle, C. Zimmer, J. Kirschke, B. Menze, et al. "Uncertainty-aware and lesion-specific image synthesis in multiple sclerosis magnetic resonance imaging: a multicentric validation study". In: ***Frontiers in Neuroscience*** 16 (2022).

Work on radiomics analysis

- **H. Li**, F.-F. Xue, K. Chaitanya, S. Luo, I. Ezhov, B. Wiestler, J. Zhang, and B. Menze. "Imbalance-aware self-supervised learning for 3d radiomic representations". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (**MICCAI**)*. Springer. 2021, pp. 36–46.

Besides the work listed above, further *first-author* publications on image analysis are listed below but are **not** included in this dissertation. These works extend the application domains and enhance the methodology.

Further own work on image analysis

- **H. Li**, R. G. Prasad, A. Sekuboyina, C. Niu, S. Bai, W. Hemmert, and B. Menze. "Micro-Ct synthesis and Inner Ear Super Resolution via Generative Adversarial Networks and Bayesian Inference". In: *2021 IEEE 18th International Symposium on Biomedical Imaging (**ISBI**)*. IEEE. 2021, pp. 1500–1504.
- **H. Li**, S. Gopal, A. Sekuboyina, J. Zhang, C. Niu, C. Pirkl, J. Kirschke, B. Wiestler, and B. Menze. "Unpaired MR Image Homogenisation by Disentangled Representations and Its Uncertainty". In: *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging (**UNSURE**), and Perinatal Imaging, Placental and Preterm Image Analysis*. Springer, 2021, pp. 44–53.
- **H. Li**, M. Reichert, K. Lin, N. Tselousov, R. Braren, D. Fu, R. Schmid, J. Li, B. Menze, and K. Shi. "Differential diagnosis for pancreatic cysts in CT scans using densely-connected convolutional networks". In: *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (**EMBC**)*. IEEE. 2019, pp. 2095–2098.
- **H. Li**, J. Zhang, and B. Menze. "Generalisable cardiac structure segmentation via attentional and stacked image adaptation". In: *International Workshop on Statistical Atlases and Computational Models of the Heart (**STACOM**)*. Springer. 2020, pp. 297–304.
- X. Dong*, **H. Li***, Z. Jiang, T. Grünleitner, İ. Güler, J. Dong, K. Wang, M. H. Köhler, M. Jakobi, B. H. Menze, et al. "3D Deep Learning Enables Accurate Layer Mapping of 2D Materials". In: ***ACS Nano*** 15.2 (2021), pp. 3139–3151.
- X. Dong*, Y. Zhang, **H. Li***, Y. Yan, J. Li, J. Song, K. Wang, M. Jakobi, A. K. Yetisen, and A. W. Koch. "Microscopic Image Deblurring by a Generative Adversarial Network for 2D Nanomaterials: Implications for Wafer-Scale Semiconductor Characterization". In: ***ACS Applied Nano Materials*** (2022).
- X. Dong*, **H. Li***, Y. Yan, H. Cheng, H. X. Zhang, Y. Zhang, T. D. Le, K. Wang, J. Dong, M. Jakobi, et al. "Deep-Learning-Based Microscopic Imagery Classification, Segmentation, and Detection for the Identification of 2D Semiconductors". In: ***Advanced Theory and Simulations*** (2022), p. 2200140.
- **H. Li**, A. Zhygallo, and B. Menze. "Automatic brain structures segmentation using deep residual dilated U-Net". In: *International MICCAI Brainlesion Workshop (**BrainLes**)*. Springer. 2018, pp. 385–393.
- **H. Li**, J. Zhang, M. Muehlau, J. Kirschke, and B. Menze. "Multi-scale convolutional-stack aggregation for robust white matter hyperintensities segmentation". In: *International MICCAI Brainlesion Workshop (**BrainLes**)*. Springer. 2018, pp. 199–207.

In addition to my own works listed above, co-authored publications in image computing, including challenge benchmarks, neuroscience, and shape modeling, are listed below. However, they are **not** included in this dissertation.

Other publications on image analysis

- V. M. Campello, P. Gkontra, C. Izquierdo, C. Martin-Isla, A. Sojoudi, P. M. Full, K. Maier-Hein, Y. Zhang, Z. He, J. Ma, H. Li, et al. "Multi-centre, multi-vendor and multi-disease cardiac segmentation: the M&Ms challenge". In: *IEEE Transactions on Medical Imaging (**TMI**)* 40.12 (2021), pp. 3543–3554.
- J. Chen, W. Li, **H. Li**, and J. Zhang. "Deep class-specific affinity-guided convolutional network for multimodal unpaired image segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (**MICCAI**)*. Springer. 2020, pp. 187–196.
- O. Schoppe, C. Pan, J. Coronel, H. Mai, Z. Rong, M. I. Todorov, A. Müskes, F. Navarro, **H. Li**, A. Ertürk, et al. "Deep learning-enabled multi-organ segmentation in whole-body mouse scans". In: ***Nature Communications*** 11.1 (2020), pp. 1–14.
- Y. Zhao, Y. Liu, Y. Kan, A. Sekuboyina, D. Waldmannstetter, **H. Li**, X. Hu, X. Zhao, K. Shi, and B. Menze. "Spatial-Frequency Non-local Convolutional LSTM Network for pRCC Classification". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (**MICCAI**)*. Springer. 2019, pp. 22–30.
- H. J. Kuijf, J. M. Biesbroek, J. De Bresser, R. Heinen, S. Andermatt, M. Bento, M. Berseth, M. Belyaev, M. J. Cardoso, A. Casamitjana, **H. Li**, et al. "Standardized assessment of automatic segmentation of white matter hyperintensities and results of the WMH segmentation challenge". In: *IEEE transactions on medical imaging (**TMI**)* 38.11 (2019), pp. 2556–2568.
- P. Carolin, M. Cencini, J. W. Kurzawski, D. Waldmannstetter, **H. Li**, and M. I. Menzel. "Learning residual motion correction for fast and robust 3D multiparametric MRI". In: ***Medical Image Analysis*** 2022.77 (2022).
- I. Ezhov, T. Mot, S. Shit, J. Lipkova, J. C. Paetzold, F. Kofler, C. Pellegrini, M. Kollovieh, F. Navarro, **H. Li**, et al. "Geometry-aware neural solver for fast Bayesian calibration of brain tumor models". In: *IEEE Transactions on Medical Imaging (**TMI**)* 41.5 (2021), pp. 1269–1278.
- A. Sekuboyina, M. E. Husseini, A. Bayat, M. Löffler, H. Liebl, **H. Li**, G. Tetteh, J. Kukačka, C. Payer, D. Štern, et al. "VerSe: A Vertebrae labelling and segmentation benchmark for multi-detector CT images". In: ***Medical Image Analysis*** 73 (2021), p. 102166.
- M. F. Thomas, F. Kofler, L. Grundl, T. Finck, **H. Li**, C. Zimmer, B. Menze, and B. Wiestler. "Improving Automated Glioma Segmentation in Routine Clinical Use Through Artificial Intelligence-Based Replacement of Missing Sequences With Synthetic Magnetic Resonance Imaging Scans". In: ***Investigative Radiology*** 57.3 (2022), pp. 187–193.
- X. Hu, Y. Yan, W. Ren, **H. Li**, A. Bayat, Y. Zhao, and B. Menze. "Feedback Graph Attention Convolutional Network for MR Images Enhancement by Exploring Self-Similarity Features". In: *Medical Imaging with Deep Learning (**MIDL**)*. PMLR. 2021, pp. 327–337.
- J. Ma, X. Li, **H. Li**, R. Wang, B. Menze, and W.-S. Zheng. "Cross-view relation networks for mammogram mass detection". In: *2020 25th International Conference on Pattern Recognition (**ICPR**)*. IEEE. 2021, pp. 8632–8638.

- X. Hu, R. Guo, J. Chen, **H. Li**, D. Waldmannstetter, Y. Zhao, B. Li, K. Shi, and B. Menze. "Coarse-to-fine adversarial networks and zone-based uncertainty analysis for NK/T-cell lymphoma segmentation in CT/PET images". In: *IEEE journal of Biomedical and Health Informatics (**MIDL**)* 24.9 (2020), pp. 2599–2608.
- K. M. Timmins, I. C. van der Schaaf, E. Bennink, Y. M. Ruigrok, X. An, M. Baumgartner, P. Bourdon, R. De Feo, T. Di Noto, F. Dubost, **H. Li**, et al. "Comparing methods of detecting and segmenting unruptured intracranial aneurysms on TOF-MRAS: The ADAM challenge". In: ***NeuroImage*** 238 (2021), p. 118216.
- D. M. Hedderich, A. Menegaux, **H. Li**, B. Schmitz-Koep, P. Stämpfli, J. G. Bäuml, M. T. Berndt, F. J. Bäuerlein, M. J. Grothe, M. Dyrba, et al. "Aberrant claustrum microstructure in humans after premature birth". In: ***Cerebral Cortex*** 31.12 (2021), pp. 5549–5559.
- A. B. Qasim, I. Ezhov, S. Shit, O. Schoppe, J. C. Paetzold, A. Sekuboyina, F. Kofler, J. Lipkova, **H. Li**, and B. Menze. "Red-GAN: Attacking class imbalance via conditioned generation. Yet another medical imaging perspective." In: *Medical Imaging with Deep Learning (**MIDL**)*. PMLR. 2020, pp. 655–668.
- Y. Zhao, **H. Li**, S. Wan, A. Sekuboyina, X. Hu, G. Tetteh, M. Piraud, and B. Menze. "Knowledge-aided convolutional neural network for small organ segmentation". In: *IEEE journal of Biomedical and Health Informatics (**JBHI**)* 23.4 (2019), pp. 1363–1373.
- Y. Zhao, **H. Li**, R. Zhou, G. Tetteh, M. Niethammer, and B. H. Menze. "Automatic multi-atlas segmentation for abdominal images using template construction and robust principal component analysis". In: *2018 24th International Conference on Pattern Recognition (**ICPR**)*. IEEE. 2018, pp. 3880–3885.
- M. Kober-Hasslacher, H. Oh-Strauß, D. Kumar, V. Soberon, C. Diehl, M. Lech, T. Engleitner, E. Katab, V. Fernández-Sáiz, G. Piontek, **H. Li**, et al. "c-Rel gain in B cells drives germinal center reactions and autoantibody production". In: ***The Journal of Clinical Investigation*** 130.6 (2020), pp. 3270–3286.
- L. Xu, G. Tetteh, J. Lipkova, Y. Zhao, **H. Li**, P. Christ, M. Piraud, A. Buck, K. Shi, and B. H. Menze. "Automated whole-body bone lesion detection for multiple myeloma on 68Ga-pentixafor PET/CT imaging using deep learning methods". In: ***Contrast media & molecular imaging*** 2018 (2018).
- C. Niu, Y. Wang, A. D. Cohen, X. Liu, **H. Li**, P. Lin, Z. Chen, Z. Min, W. Li, X. Ling, et al. "Machine learning may predict individual hand motor activation from resting-state fMRI in patients with brain tumors in perirolandic cortex". In: ***European Radiology*** 31.7 (2021), pp. 5253–5262.
- G. Al Boustani, L. J. K. Weiß, **H. Li**, S. M. Meyer, L. Hiendlmeier, P. Rinklin, B. Menze, W. Hemmert, and B. Wolfrum. "Influence of Auditory Cues on the Neuronal Response to Naturalistic Visual Stimuli in a Virtual Reality Setting". In: ***Frontiers in Human Neuroscience*** 16 (2022).

1 Introduction

Neuroimage analysis has made fundamental contributions to the study of neuroscience [1, 2], image-guided surgery [3, 4], and neurological diagnostic tasks [5, 6]. Various imaging techniques have been developed to generate visual contrast of brain structures, such as multi-contrast magnetic resonance imaging (MRI) [7, 8], positron emission tomography [9, 10], and computed tomography [11, 12]. Particularly, MRI provides high-resolution contrast for soft brain tissues. For example, T1-weighted sequences can distinctly differentiate the gray and white matter of the brain [13] and reflect the degree of brain atrophy. In practice, multi-contrast neuroimages offer complementary structural information, which could be leveraged to identify sub-structures, e.g., in multiple sclerosis [14, 15] and brain tumor cases [16, 17].

However, interpreting multi-contrast neuroimages is a complex task that necessitates domain knowledge and is predominantly performed by neuro-radiologists. The benefits of applying machine learning, particularly deep learning, in neuroimage analysis are threefold: (1) reducing expert effort, thus accelerating neuroimage analysis from small-scale to population-based studies, (2) augmenting image information from hardly-visible to visible, thereby improving clinical diagnosis, and (3) discovering new meaningful features or patterns from large-scale datasets, which can be either labeled or unlabeled.

Deep learning [18] has become ubiquitous in neuroimage analysis. With the increase in neuroimaging data volume and advances in hardware, these data-driven 'black-box' algorithms have enabled substantial leaps in performance across a wide array of neuroimage analysis applications [19, 20]. However, deploying existing deep learning models in clinics and laboratories remains a challenge. A combination of labeling scarcity, missing data, and domain shift [21] between image datasets forms a high barrier to adoption and results in underwhelming performance in real-world settings. How deep learning can be efficiently adopted, e.g., in scenarios with limited or no annotations available, and how techniques can be adapted to real-world applications under domain shifts, are both technically attractive and clinically relevant. Notably, such developed techniques should be carefully evaluated and interpreted in clinical settings.

This dissertation focuses on three advanced neuroimage analysis tasks: image segmentation, image synthesis, and radiomics analysis. It develops and discusses several machine learning techniques in the context of neuroimage analysis, including deep transfer learning, generative modeling, domain adaptation, and contrastive learning. Specifically, it addresses the following research questions to advance neuroimage analysis:

- How to develop effective and efficient deep learning approaches for brain structure segmentation in MR images? How can the prior knowledge learned from one task in one image domain be transferred to similar tasks in different image domains?

- Can one synthesize a missing contrast image (i.e., from multiple source image domains to a target image domain) in multi-contrast MRI? If yes, how realistic and useful are synthetic data in clinical practice?

- How to improve pathology-specific synthesis quality and the robustness of generative adversarial networks in the presence of domain shifts?

- Can self-supervised learning-based representation serve as a new kind of radiomics? If yes, is there any potential imbalance issue in the learned representation?

1.1 Organization

This publication-based dissertation is structured as follows:

Chapter 1 introduces three neuroimage analysis tasks: (a) image segmentation, (b) image synthesis, and (c) radiomics analysis, along with existing works and challenges. It concludes with a summary of the **dissertation's contributions**. Chapter 2 provides an overview of relevant terminology and key concepts used throughout this manuscript.

Chapter 3 to 7 are composed of eight publications in their original forms. They have been published as peer-reviewed journals and conference proceedings and are therefore self-contained. Each of these chapters begins with a summary that includes the full citation of the original publication, a short synopsis introducing the content of the corresponding publication, and the author's contributions.

Chapter 8 offers a discussion and conclusions over the presented material and suggests directions for future work.

1.2 Medical image segmentation

1.2.1 Motivation and challenges

Image segmentation is one of the main challenges in modern medical image analysis and the base of many quantitative image analysis tasks. It aims to label an image's pixels (or voxels) into multiple semantic regions for localization and quantification purposes. The outcome of semantic image segmentation can be a set of anatomical labels or contours that highlight regions of interest. Since manual segmentation of target structures is labor-intensive and requires expert knowledge [22], automated image segmentation has become one of the main research topics in medical imaging. It benefits many medical applications such as computer-aided diagnosis, therapy planning, lesion detection, and disease progression [23, 16, 24, 25, 26, 27]. Although various segmentation approaches have been developed and reported, the task remains challenging at diverse clinical and technical settings. It seems that there is no general solution given the following obstacles:

- Labeling uncertainty: high inter-rater variability exists for many segmentation tasks, especially the segmentation of small structures (e.g., multiple sclerosis lesions). Such uncertainty is caused by two main sources: (a) poor image quality, such as low contrast between structures and the noise that corrupts image structure, and (b) different clinical opinions on the same structure.

- Inhomogeneity of individuals: Anatomical structures are often inhomogeneous among healthy subjects and patients, concerning the shape of organs and the texture of pathology. It may degrade the segmentation performance since the training set can normally only cover a part of underlying distributions.

- Domain shift in image acquisitions: practically, collecting sufficient training data with the same acquisition setting as the future ones during the inference stage is impossible. In MR imaging, such a domain gap relates to bias field, image resolution, and image contrast

- Partial and missing annotations: existing segmentation pipelines only apply for offline data (with fixed target structures). For new applications, e.g., one wishes to segment new anatomical structures, existing annotations cannot be straightforwardly used for training or updating new models. Furthermore, different data subsets may have varied partial annotations.

1.2.2 State-of-the-art methods

Previous classical approaches should be appreciated, such as rule-based [28], statistical inference based [23, 29, 23, 30], level set based [31, 32] and atlas-based [33, 34] methods. Due to topic relevance, only two groups of recent methods are reviewed in detail: (1) machine-learning-based segmentation with hand-crafted features, and (2) deep-learning-based segmentation.

Hand-crafted features and machine-learning-based segmentation

- Unsupervised methods: clustering methods include k-means [35] and fuzzy c-means clustering [36] with segmentation refinement by deformable models [37] are often used to group the pixels into multiple classes. They divide the features of data into two or more clusters and assign the identities to each data point (i.e., pixels). The k-means method generates results corresponding to hard segmentation while fuzzy c-means produces soft segmentation that can be transformed into hard segmentation by allowing the pixels to have cluster identity.

- Supervised methods: in this group, segmentation is formulated as an optimization procedure to learn a decision boundary that best separates binary classes. They leverage hand-crafted features including intensity [38], texture [39, 40], and shape [41] features together with machine learning techniques such as support vector machine [42], random forest [43], vantage point forest [44], dictionary learning [45] to classify each pixel to a semantic label. However, these methods rely on careful pre-processing, and feature engineering and are commonly multi-stage.

Deep-learning based segmentation

- Architectures: deep neural architectures, also called 'backbone', containing most parameters, is optimized to learn a mapping from image space to label space. Kamnitsas *et al.* [46] is the first to develop a 3D encoder-like convolutional neural network for automated brain lesion segmentation in MR images. Although it improved segmentation accuracy compared to traditional machine learning approaches, the computation of the inference stage is relatively heavy because it relies on sliding windows to obtain the labels of all voxels in a volume. Fully convolutional network (FCN) [47] and U-Net, which are encoder-decoder-like architectures, are proposed to perform pixels-to-pixels *dense* prediction. Such fully convolutional architectures can take arbitrary image sizes as the input and generate the corresponding segmentation maps at once. Based on U-Net, many variants, such as U-Net++ [48], are proposed to improve the effectiveness of representation learning or tailor it for specific application requirements. Neural architecture search is an emerging direction for optimizing architecture itself [49]. An existing framework named nn-UNet is rule-based with a few searching parameters (e.g., input size and network complexity). It is well validated on diverse imaging datasets and proven to be practically effective [50].

- Loss functions: the role of a loss function is to evaluate how well the predicted segmentation matches the reference (or ground truth). Various loss functions have been proposed for medical image segmentation. Cross entropy (CE) is derived from *Kullback-Leibler* divergence, which is a measure of dissimilarity between two distributions. Focal loss [51] modifies the standard CE to handle extreme foreground-background imbalance issues. Dice loss [52] can directly optimize a Dice coefficient and handle class imbalance issues. Tversky loss [53] extends Dice loss and emphasizes reducing false negatives. Distance-based loss [54] is orthogonal to the above ones and aims to minimize the distance metric between the prediction and the reference. Another type of loss function is feature-based instead of pixel-wise. Recently, a perceptual loss has been shown to be useful in image segmentation [55]. Considering the complementary properties of different loss functions, one could combine multiple losses for specific segmentation tasks. For example, a weighted sum between Dice and distance-based losses can improve the segmentation of boundary pixels [54].

- Domain adaptation and generalization: how much data for a segmentation network to learn and perform well on unseen distribution is related to the concept of domain adaptation. Given an unseen target domain normally different from the source domain where the model is trained, adapting deep neural networks can be partially mitigated with techniques such as transfer learning [56, 57]. It uses small labeled datasets from the target domain to fine-tune the segmentation model trained on the source domain. However, in this scenario, annotations from the target domain are needed. Instead of using supervised learning, another way to perform adaptation is unsupervised, i.e., using only images and without using labels from the target domain. Kamnitsas *et al.* [58] is the first to develop unsupervised domain adaptation techniques for brain lesion segmentation using adversarial training. In a more practical setting, Karani *et al.* [59] proposes test-time adaptation for medical image segmentation where only one data sample is used during the adaptation process. Cross-modality adaptation [60, 61, 62] is a very challenging problem given a

large domain gap between the source and the target, e.g., from CT to MRI, although it is debatable if this setting is clinically relevant. Domain adaptation methods discussed above may not always be practical as they rely on training data from the target domain and a fine-tuning stage, hindering their deployment in clinical practice. Domain generalization aims to improve the robustness of out-of-distribution data. In particular, single-source domain generalization [63, 64] tacks with a practical scenario in which the model is optimized on data from only one source domain.

- Pre-training: Instead of training the model from scratch, pre-training the neural architecture is proven to improve the effectiveness of representations. Unlike transfer learning or domain adaptation discussed above, pre-training allows the model to explore the training set's data structures without additional supervision. Random initialization or initialization with specific distributions [65, 66, 67] is proven to improve the training of neural networks. Pre-training with pre-text tasks is a self-supervised approach to initialize the weights and obtain a representation. Zhou *et al.* [68] systematically demonstrate that pre-training on medical data with self-supervision consistently outperforms pre-training on ImageNet [69]. Contrastive learning-based pre-training is another effective self-supervised method by learning invariance from data augmentations. Chaitany *et al.* [70] propose a multi-scale contrastive learning loss for 3D medical data and demonstrate its effectiveness for medical image segmentation tasks.

1.3 Medical image synthesis

1.3.1 Multi-contrast MRI

Multi-contrast MR imaging enables neuro-anatomies to be imaged under different contrasts by manipulating pulse MR sequences [71]. Hence, neuroimages acquired from several distinct contrasts can help better distinguish tissues and enhance diagnostic information than single contrast alone. For example, gray and white matter can be better visually separated in T1-weighted images, whereas white matter hyperintensities can be well detected in FLAIR images [72]. The limitations of multi-contrast acquisition are its long scan time and excessive artifacts caused by motion or unexpected signal [73]. Therefore, neuroimage synthesis of missing or corrupted contrasts from other high-quality contrasts is promising to improve the clinical utility of multi-contrast MRI.

1.3.2 Prior works

Prior image synthesis approaches for multi-contrast MRI can be categorized into two main streams: one-to-one and many-to-one methods.

- One-to-one synthesis: In one-to-one synthesis, the objective is to generate a subject's image in a target contrast from the same subject's image in a source contrast. Early research formulated one-to-one synthesis as a sparse dictionary learning problem [74, 75] where patch-based dictionaries are formed from a set of co-registered source atlas and target atlas. Each patch in the source is represented as a sparse linear combination of dictionary atoms of the source atlas, and this combination is then used to synthesize a target contrast image from the target atlas. To improve the synthesis quality, patch-based non-linear regression with random forests [76] has been used for the source-to-target mapping. To overcome the limitation of patch-based approaches, a deep convolutional neural network (CNN) was proposed to process the entire 2D slices [77]. Recent approaches further leveraged adversarial training and additional feature-based supervision to better capture the local details in the target domain [78, 79].

- Many-to-one synthesis: when several source contrasts are available, to better leverage multiple sources is to perform many-to-one synthesis. It aims to generate one subject's image in the target contrast from several source images in different contrasts. One popular approach is to learn a non-linear regression model using random forests [80, 81]. Random forests fit a regression model in the feature space (e.g., texture feature) to estimate the intensities of the target contrast given multiple source contrasts. Similar to one-to-one methods, CNN has been developed for many-to-one synthesis [82, 83], and recent works have employed an adversarial loss to improve the image quality [84, 85].

In general, one-to-one synthesis aims to generate the target image from the feature representation of a given source image. Since it is optimized for a single input contrast, such a mapping can be effectively learned when the source and target contrasts are highly *correlated*, and it might limit synthesis quality when two contrasts are weakly linked. For example, beyond the multi-contrast MRI domain, CT and MRI are highly complementary imaging tools to visualize hard and soft tissues, respectively. Thus, a correct mapping from CT to MRI might be challenging to learn as the contrast of soft tissues is not well presented in CT images. On the other hand, many-to-one synthesis attempts to recover the target image from a shared representation of multiple sources. It naturally manifests enhanced diagnostic information that is shared across distinct source contrasts, even when this information is weakly shown in individual contrasts.

1.4 Radiomics analysis

1.4.1 Definition and pipeline

In the cancer field, simple measurements of tumor (e.g., tumor size [86]) do not reflect the morphological complexity or behavior [86]. They do not cover the range of quantitative features such as edge, texture, or shape. Hence, radiomics refers to the high-throughput extraction and analysis of advanced quantitative features from radiological images, such as CT, PET, and MRI. Such profound analysis and feature mining can uncover predictive or prognostic correlations between images and clinical outcomes.

The radiomics pipeline can be divided into five standard steps with definable inputs and outputs: (a) image acquisition, (b) image segmentation and rendering, (c) feature extraction and quantification, (d) data sharing, and (e) downstream task analyses. Notably, there are variations and challenges from any of these individual processes. For example, different image acquisition settings might produce different sets of features for the same region of interest. Thus, after optimization, one existing challenge is to stabilize the entire process [87]. We concentrate on radiomic features and review them in detail in the following.

1.4.2 Existing radiomic features

Once a region of interest (e.g., lesion) is defined, we can extract quantitative image features which can be grouped into two categories:

- Rule-based features: these features describe characteristics of the intensity histogram, shape, texture, as well as spatial descriptors of ROI location and relations with the surrounding structures. Tumor intensity histogram-based features reduce the data dimension from a volume into a single histogram. It describes the fractional volume for a selected structure given a range of voxel values [88, 89]. Shape features that describe the geometric shape can be extracted from the 3D surface of the rendered volumes [90], e.g., tumor compactness and sphericity [91]. Texture features, including co-occurrence features and local descriptors, are widely applied in medical classification tasks [92, 93, 94, 95]. The basis of the co-occurrence features lies in the second-order joint conditional probability density function of the given texture. The elements of the co-occurrence matrix for the ROI represent the number of times that two intensity levels occur in two voxels separated by the distance in the direction. Subsequently, features can be extracted from this conditional probability density function, e.g., describing contrast, correlation, cluster prominence, cluster shade, cluster tendency, maximum probability, dissimilarity, energy, homogeneity, a sum of squares, sum average or sum variance, etc. Local descriptors such as local binary patterns [96] and Gabor filters [97] are used, along with feature coding methods to enhance the feature representation, such as a bag of words [98], and Fisher vector [99].

- Data-driven features: Instead of extracting pre-defined rule-based features, deep learning methods can uncover high-level abstract information from raw imaging data as the parameters are optimized based on tasks in an end-to-end manner. These deep features may provide more predictive power than rule-based ones, although they suffer from an increased training data size and weaker interpretability issues. For instance, the deep features extracted from a CNN model can predict lung cancer survival [100]. Other examples are that such features can be used to discriminate benign and malignant lung tumors [101], and to extract semantic features from breast mammography images [102]. They are proven to reach expert-level performance in some considerably challenging applications [103, 104]. However, data-driven radiomics features also require addressing new challenges and facing several issues. These include the need for appropriate training with prior knowledge due to the limited size of available datasets and the high level of heterogeneity, especially when

training networks from scratch and facing domain shifts. Recent studies show that some success can be achieved with transfer learning from medical images (e.g. *Rad-ImageNet* [105]) and reducing the amount of training data. Instead of adopting supervised learning, unsupervised (or self-supervised) approaches such as auto-encoder can compress the raw image space into a low-dimensional latent space without clinical labels [106].

1.5 Summary of contributions

This dissertation covers the introduced three neuroimage analysis tasks: image segmentation, image synthesis, and radiomics analysis. In the following, a brief introduction to the motivation and contribution of each publication-based chapter is presented.

Chapter 3: Effective brain lesion and claustrum segmentation with ensembles of deep nets

This chapter establishes a deep-learning, ensemble-model-based framework for effectively segmenting brain structures in MR images. Two distinct tasks - white matter hyperintensities segmentation and claustrum segmentation are investigated.

Motivation. White matter hyperintensities (WMH) are commonly observed in FLAIR MRI scans of elderly people, and they are associated with many neurological disorders such as cognitive decline and dementia [107, 108, 109]. Manual pixel-wise delineation of WMH regions is a subjective way to assess abnormalities and disease progression. However, this procedure is labor-intensive and nonautomated for neuroradiologists. It shows high intra-rater, and inter-rater variability [110]. The WMH Segmentation Challenge 2017 [111] was held to benchmark segmentation algorithms at the 20^{th} International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI, 2017). We submitted a U-Net and ensemble model-based solution to this challenge, which was containerized and evaluated by the organizers on clinical multi-center datasets. The final test set includes 110 hidden cases from five different MR scanners from three hospitals in the Netherlands and Singapore. Notably, the test sets remained unseen to our team and other participants.

The claustrum is a thin and sheet-like gray matter structure of the mammalian forebrain between the striatum and insular cortex, and it is connected to both the ipsilateral and the contralateral hemisphere [112, 113, 114]. Recent studies suggest that it plays a role in fundamental cognitive processes such as selective attention or task switching [115, 116]. Another interesting perspective on the claustrum is its unique ontogeny and link to subplate neurons, which relates to neurodevelopmental disorders such as preterm birth [117, 118]. Similar to white matter lesions, claustrum segmentation is notoriously time-consuming, requires expert knowledge of the brain anatomy, and is not feasible in large-scale studies of the human brain [119, 120]. Thus, to advance our understanding of the claustrum in humans in population studies, an objective, accurate, automated, and robust MRI-based segmentation method is in need.

Contributions and results. Our approach to segment WMH in MR images is based on an ensemble of U-shape architectures with long-range connections [121]. We optimized individual models with random weight initialization shuffled batches and aggregated them to vote the pixel labels n. In the evaluation stage, the containerized method was submitted to the challenge organizers, who then independently tested it on hidden one hundred and ten cases acquired from five MRI scanners. We achieved the best Dice score, lesion-level precision, and robust Hausdorff distance of 80%, 84%, and 6.30 mm, respectively, on held-out test datasets. These numbers suggest that our method is state-of-the-art. We present a cross-scanner study to discuss how the combination of modalities impacts on generalizability. Importantly, we presented a quantitative study to show the effect of ensemble size and the effectiveness of the ensemble model. We demonstrate that the proposed method has the potential to be applied in clinical practice.

For claustrum segmentation, we develop a multi-view ensemble-based approach to segment the adult claustrum in T1-weighted MRI scans. We train two single-view models on the 2D slices from axial and coronal views, respectively. During the inference stage, we predict the single-view segmentation masks from two views and aggregate them by averaging the voxel-wise probabilities. We trained and evaluate our method in 181 individuals, which were manually

annotated by a neuroradiologist as a reference. We achieve a median volumetric similarity, a robust Hausdorff distance, and a Dice score of 93.3%, 1.41 mm, and 71.8%, respectively, representing equal or superior segmentation performance. In a leave-one-scanner-out evaluation, we demonstrate good model generalizability on the images from unseen scanners at slightly inferior accuracy. Furthermore, we find that our segmentation approach benefits from multi-view aggregation. We conclude that the proposed method enables fast and robust automated claustrum segmentation and thus yields considerable potential for accelerating MRI-based research of the human claustrum.

Chapter 4: Efficient neonate claustrum segmentation with deep transfer learning

The chapter explores deep transfer learning to reduce annotation effort significantly. The research questions are: given the same segmentation task but in a different image domain, do we need to annotate a large amount of data from scratch as we did in the previous chapter? Or can we leverage prior knowledge to reduce further annotation efforts?

Motivation. Claustrum development in humans is not well explored in neuroscience and relies on MRI studies. Most studies focus on animals, while macrostructural and microstructural maturation in humans remain unknown [122, 123]. Hence, close examination and characterization of claustrum development in younger cohorts are of special interest; however, MRI data about the claustrum in a sizable neonatal cohort are missing, mostly due to the lack of adequate automated segmentation methods. Recently, automated segmentation of the human claustrum in adults has been investigated by structural approximation to the dorsal claustrum [124] and a 2D deep-learning approach [125]. Furthermore, in the previous chapter, a multi-view deep learning-based model has been proposed [126] to segment the adult claustrum trained on a large annotated dataset; however, there is no reliable automated segmentation method available for the claustrum in neonatal MRI. It might be sub-optimal to perform manual segmentation of neonate claustrum in 100+ T2-w scans from scratch and then using them to re-train the model.

Contributions and results. We transfer the knowledge learned from *adult* claustrum segmentation in T1-weighted scans to the new application. We develop a deep transfer learning-based method to segment the claustrum in 558 T2-weighted neonatal brain MRI of the developing Human Connectome Project (dHCP). We train and evaluate the method on 30 manual bilateral claustrum annotations in neonates. With only *twenty* labeled scans, the method achieves a median volumetric similarity, a robust Hausdorff distance, and a Dice score of 95.9%, 1.12 mm, and 80.0%, respectively, representing an excellent agreement between the automated segmentation and the reference. When comparing with inter-rater reliability, the method achieves significantly superior volumetric similarity ($p = 0.047$) and Dice score ($p < 0.005$). Furthermore, we demonstrate the effectiveness of deep transfer learning compared with non-transfer learning. We observe that the model achieved satisfactory segmentation with only *twelve* annotated scans. Finally, we confirm the model's applicability on 528 scans and reveal its reliable segmentation in 97.4%.

Chapter 5: Unified multi-contrast MR neuroimage synthesis and its clinical validation

This chapter develops a novel generative adversarial network to perform *multiple-to-one* cross-modality synthesis for multi-contrast MRI and validates the synthetic images in a clinical study.

Motivation. MRI datasets often consist of high-dimensional image volumes, multiple contrasts, and repeated scans acquired at multiple time points. The imaging protocols broadly vary depend-

ing on the imaging centers, which hinders the comparability between one another. Particularly in multiple sclerosis brain imaging, double inversion recovery (DIR) is a sensitive sequence to detect lesions but might not be a standard imaging sequence for all hospitals. However, existing lesion quantification tools require identical modalities at multiple time points. Another example is glioma imaging. Although automated glioma segmentation [16] holds promise for objective assessment of tumor response, its routine clinical use is impaired by missing sequences due to motion artifacts.

Potentially, *cross-modality image synthesis* can resolve those obstacles through data infilling and re-synthesis. In multi-contrast MRI, *multiple-to-one* cross-modality mapping is highly relevant as proprietary information of distinct individual modalities can be synergistic. There are several challenges in cross-modality medical image synthesis: (1) the input and target modalities are assumed to be not spatially aligned because registration methods for aligning modalities may fail. It hinders the applicability of conventional regression methods. (2) input modalities may be incomplete due to different clinical settings; thus, a traditional regression-based approach would be restricted to the smallest uniform subset. (3) existing methods have limited scalability. For example, one would need to train separated models for possible combinations of the input domains when using fixed input and output, such as the Cycle-GAN setting [127]. Another open question is how to evaluate the quality of synthetic images. Traditional metrics such as PSNR and structural similarity [128] can provide quantitative quality measurement, but they cannot quantify whether the image captures clinically relevant substructures. Hence, such synthetic images should be interpreted by experts who can distinguish the image quality.

Contributions and results. We propose *DiamondGAN*, a unified, scalable multi-modal generative adversarial network. It learns a multiple-to-one cross-modality mapping among non-aligned modalities using only a pair of generators and discriminators optimized with a multi-modal cycle consistency loss function. We provide qualitative and quantitative results on two clinically relevant MRI sequence synthesis tasks, showing *DiamondGAN* 's superiority over baseline models. We present the results of an extensive visual evaluation performed by fourteen experienced radiologists to confirm the quality of synthetic images. We observed that trained radiologists cannot distinguish our synthetic double inversion recovery (synthDIR) images from real ones.

In a clinical study, we generate *synthDIR* images and compare their diagnostic performance to conventional sequences in multiple sclerosis patients. These images and conventionally acquired DIR (trueDIR) and FLAIR images were assessed for MS lesions by two independent readers, blinded to the source of the DIR image. Lesion counts in the different modalities were compared using a Wilcoxon signed-rank test, and inter-rater analysis was performed. Contrast-to-noise ratios were compared for objective image quality. We observed that the utilization of *synthDIR* allowed the detection of significantly more lesions than the use of FLAIR image.

Chapter 6: Lesion-specific, uncertainty-aware, and domain-adaptive synthesis

Motivation. In the previous chapter, we demonstrated that generative adversarial networks could synthesize a target contrast MRI from multi-contrast input and improve lesion detection compared to FLAIR and T2-w sequences. Nonetheless, MS lesions typically are very small, GANs are at risk of synthesizing images of high morphologic similarity to the target image while failing to translate the clinically important lesions. The ability of a GAN to learn about pathology-specific anomalies might open the door for further customization and improvements in this regard. Such domain knowledge has improved a network's training stage in various classification tasks, such as the categorization of breast lesions and the detection of malignant thyroid nodules [129, 130]. Targeted translation of parenchymal lesions and visualization of model confidence (i.e., uncertainty quantification), can further augment their utility in practice.

In addition to lesion-specific and uncertainty-aware properties, it is an open question of how GANs can be adapted to different image domains that are non-identical to the training set. However, existing synthesis frameworks are mostly developed and evaluated on single-domain data (e.g., images from the same scanner), with limited consideration of model robustness when testing on unseen image domains that might be collected from another imaging scanner or acquisition protocols. Although we have developed an uncertainty quantification to enable us to be aware of unreliable synthetic pixels, it does not address the issue of domain shift. Hence, domain adaptation is crucial for real-world deployment in clinical routine. In particular, unsupervised domain adaptation (UDA) is more practical as it does not require additional expensive supervision to fine-tune the pre-trained model.

First, there is a technical difference in domain adaptation between image synthesis and image segmentation (as we will present in detail in the paper). Second, previous works developed 2D or patch-based adaptation for image segmentation [60, 58]. Although these works show promising results, they are limited to 2D or patch domains which are insufficient for many applications, such as neuroimaging data which requires domain adaptation in a 3D fashion. The 3D image-to-image synthesis model dealing with full-volume imaging data is heavy-weight compared to the patch-based one. However, extending existing work from 2D to 3D is non-trivial. In addition to model complexity, another challenge is that the number of 3D volumetric samples is very limited, while 2D slices are more accessible.

Contributions and results. First, we tailor a loss function to improve GANs for synthesizing high-contrast DIR images and propose using uncertainty maps to enhance its clinical utility and trustworthiness further. We train GANs to synthesize DIR from a training set of FLAIR and T1-w scans of 50 MS patients. In another 50 patients as a test set, two blinded readers independently quantified lesions in synthetic DIR (synthDIR), acquired DIR (trueDIR), and FLAIR. Of the 50 test patients, 20 are acquired on the same scanner as training data (internal data), while 30 are acquired at different scanners with heterogeneous field strengths and protocols as external data. Lesion-to-background ratios (LBR) for MS-lesions *vs.* normal appearing white matter, as well as image quality parameters, were calculated. Significantly more MS-specific lesions were found in synthDIR compared to FLAIR. Importantly, improvements in lesion counts were similar in internal and external data. Measurements of LBR confirmed that lesion-focused GAN training significantly improved lesion conspicuity. We generate uncertainty maps to visualize model confidence by Monte-Carlo dropout. We observe that the use of uncertainty maps furthermore discriminates between MS lesions and artifacts.

Second, we introduce unsupervised domain adaptation for 3D medical image synthesis and present the technical difference with the existing setup in image classification and segmentation. We propose an efficient 2D variational-autoencoder approach to perform UDA in a 3D manner. We present empirical studies on the effects of the amount of data needed for adaptation and the effect of critical hyper-parameters. We show that the proposed method can significantly improve the synthesis quality on unseen domains. The proposed 2D s-VAE method outperforms both heuristic data augmentation and a 3D VAE method in two tasks. 3D VAE encoder is more computationally expensive since the encoder of 3D VAE has 5.17 million learnable parameters while 2D s-VAE only has 1.73 million ones. Although there is still a visible performance gap between all the UDA methods and the upper bound, our 2D s-VAE method provides an effective and efficient solution when the output modality from the target domain is not accessible

Chapter 7: Imbalance-aware self-supervised radiomics

Motivation. Radiomics can quantify the characteristics of regions of interest in medical data. Classically, they are pre-defined statistics of shape, texture, and other low-level image features. As an alternative, supervised deep learning-based representations are effective task-specific features but require expensive annotations and often suffer from over-fitting and data imbalance issues.

Recent contrastive learning-based methods [131, 132, 133] learn informative representations without supervision from manual labeling. However, they often rely on large training batches, and most of them only work for 2D images. Due to the high dimensionality and the limited number of training samples in the medical field, applying contrastive-learning-based approaches might not be practically feasible in a 3D setting. In this study, we identify two main differences required to adapt such self-supervised representation learning for the medical domain compared to the natural image domain: i) Medical images are often multi-modal and three-dimensional. Thus, learning 3D representation for medical imaging would be computationally costly. ii) medical datasets are heterogeneous and inherently imbalanced, e.g., distribution disparity of disease phenotypes. Existing methods are built upon approximately balanced datasets (e.g., CIFAR [134], and ImageNet [69]) and do not assume the existence of data imbalance. Thus, handling data imbalance is not well investigated in the context of self-supervised learning.

Contributions and results. We address the challenge of learning an effective representation of a 3D medical image without supervision under data imbalance. We propose a self-supervised representation learning framework to learn high-level features of 3D volumes to complement existing radiomics features. Specifically, we demonstrate how to learn image representations in a self-supervised fashion using a 3D Siamese network. More importantly, we deal with data imbalance by exploiting two unsupervised strategies: a) sample re-weighting and b) balancing the composition of training batches. We observe that combining the learned self-supervised feature with traditional radiomics improves brain tumor classification and lung cancer staging tasks covering MRI and CT imaging modalities.

2 Background

This dissertation covers several deep learning-related topics, including deep convolutional networks, generative adversarial networks, transfer learning, domain adaptation, and contrastive learning. Hence, this chapter aims to present key concepts and formulations used throughout the dissertation, but it is not intended to be a comprehensive overview of each topic. For complete and in-depth theory and analysis, please refer to book sources [135, 136, 137, 138].

2.1 Neural network

Neural networks are non-linear statistical modeling to model complex relationships between inputs and outputs and to find patterns in data. A neural network contains a number of connected units called *neurons* and is organized in layers. One example is Fig. 2.1 which contains an input layer where data is fed to the network, three hidden layers to process the data as it passes through, and an output layer to generate a prediction. The network is optimized to recognize informative patterns in the training data by comparing the output, and the actual reference under an objective function [22]. Each neuron in the network can be formulated as follows:

$$\hat{f}(\mathbf{x}) = h(\mathbf{w}^{\mathbf{T}}\mathbf{x} + b) \tag{2.1}$$

where $\mathbf{x}$ denotes the input vector, $\mathbf{w} = (w_1, \cdots, w_n)$ is the weight vector and b is called *bias* to shift the sum of weighted input by adding a constant. $h(\cdot)$ is the activation function to transform the input signals.

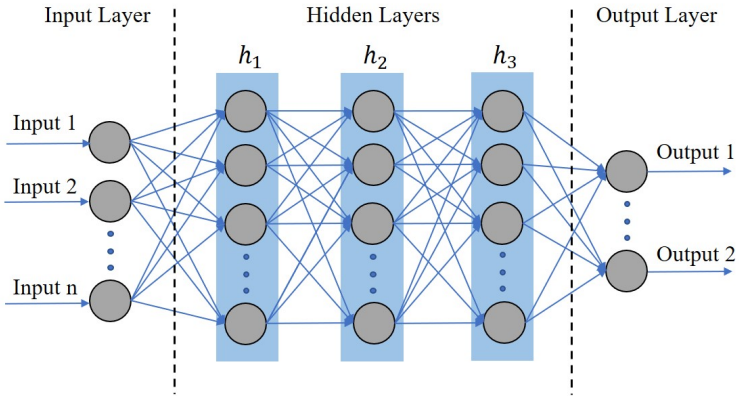


Figure 2.1: An example of neural network architecture with five layers. It consists of an input layer, three hidden layers, and an output layer. These layers are composed of a number of connected units called *neurons*. Taken from [139]

The activation functions define the output of a neuron given an input or set of inputs and are often non-linear. The commonly used ones are:

Sigmoid

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{2.2}$$

Tanh

$$\sigma(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{2.3}$$

Rectified linear unit (ReLU) [140]

$$\sigma(x) = \max(0, x) \tag{2.4}$$

Leaky ReLU [141]

$$\sigma(x) = \max(0.1x, x) \tag{2.5}$$

Maxout [142]

$$\sigma(x) = \max(w_1^{\mathbf{T}} x + b_1, w_2^{\mathbf{T}} x + b_2) \tag{2.6}$$

Swish [143]

$$\sigma(x) = \frac{x}{1 + e^{-\beta x}} \tag{2.7}$$

After combining all neurons into one layer, a network with one hidden layer can approximate any continuous function $\hat{f}(x)$ on a compact subset of $\mathbb{R}^n$, which can be formatted as a linear combination of N individual neurons:

$$\hat{f}(\mathbf{x}) = \sum_{i=0}^{N-1} a_i h(\mathbf{w}_i^{\mathbf{T}} \mathbf{x} + b_i) \tag{2.8}$$

where a_i is the weights for each neuron. All trainable parameters of the network can be summarized as:

$$\theta = (a_0, b_0, \mathbf{w_0}, \cdots, a_N, b_N, \mathbf{w_N})^{\mathbf{T}}. \tag{2.9}$$

In order to increase the model capacity and non-linearity, we can introduce more non-linear hidden layers between the input layer and output layer, as shown in Fig. 2.1. These layers are connected to a deep neural network (called a 'multilayer perceptron'). One theory on the approximation ability of neural networks is that shallow and deep networks can arbitrarily approximate any continuous function on a compact domain.

Deeper architecture benefits feature representation by reorganizing weights along different paths and re-using latent features [144, 145]. A deep neural network with several layers can be formulated as follows:

$$\begin{aligned} \hat{f}(\mathbf{x}; \mathbf{\Theta}) &= (f_m \circ \cdots \circ f_1)(\mathbf{x}) \\ &= h^m \left(h^{m-1} \left(\cdots \left(h^2 \left(h^1(\mathbf{w}_1^{\mathbf{T}} \mathbf{x} + b_1\right) + b_2\right) + b_{m-1}\right) + b_m\right) \end{aligned} \tag{2.10}$$

where $\Theta = \{\mathbf{w}_1, \cdots, \mathbf{w}_m, b_1, \cdots, b_m\}$ is the set of learnable parameters.

Parameters Θ are learned in the training phase from a training set. The training phase can be described as an optimization process of minimizing the error between the prediction and the reference. The optimization is non-linear and non-convex; hence there is no analytic solution for the parameter set Θ. Therefore, the gradient descent algorithm is utilized to learn the parameters iteratively. The back-propagation strategy [146] can efficiently compute the gradient and update the parameters Θ as:

$$\mathbf{\Theta}^{(\tau+\mathbf{1})} = \Theta^{(\tau)} - \eta \nabla \mathbb{E}(\Theta^{(\tau)}). \tag{2.11}$$

where η is the learning rate and τ denotes the iteration index.

2.2 Convolutional neural networks

Convolutional neural networks are a tailored and regularized version of deep neural networks for image-based tasks such as computer vision and medical imaging. It can capture spatial and structural information in 2D or 3D images and benefits from mechanisms including local receptive field, weight sharing, and down-sampling [147]. Different neurons' receptive fields partially overlap so that they cover the entire visual field.

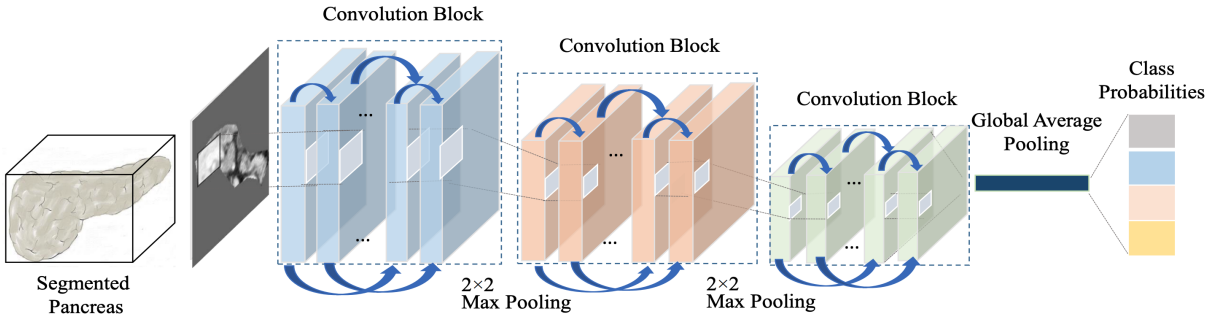


Figure 2.2: An example of convolutional neural network architecture for 3D medical image segmentation. It consists of convolutional layers and pooling layers.

According to existing designs, networks for image classification typically contain convolutional layers, pooling layers, skip connections, and fully connected layers. On top of these layers, networks for image segmentation usually contain up-sampling layers to interpolate feature maps for generating final segmentation maps with the same size as the input image. For image classification, there are milestone architectures over these years, including AlexNet [148] (2012), VGG-Net [149] (2014), Inception [150] (2015), ResNet [151] (2015), DenseNet [152] (2017), and EfficientNet [153] (2019). For image segmentation, some of the most popular architectures are fully convolutional network (FCN) [47], U-Net [121, 154], V-Net [52] and their variants [155, 156, 157, 48]. They are widely used as backbones for specific applications. Fig. 2.2 presents the structure of a convolutional neural network for image classification, and Fig. 2.3 shows the U-Net architecture for image segmentation. We introduce several functional layers in the following:

Convolutional layer. The convolutional layer convolves the input and passes its result to the next layer. It extracts local features at different locations from the previous layer (including inputs) and maps the information into a new feature map. During the convolution, the input from the previous layer is split into perceptrons, creating local receptive fields and finally compressing the perceptrons into feature maps.

At l^{th} layer, assuming a set of N^l learnable filters, each filter extracts a particular feature at certain positions on the input. The output of layer l denoted as $Y^{(l)}$ will contain N^l feature maps, where the i^{th} feature map $Y_i^{(l)}$ can be computed as:

$$Y_i^{(l)} = h\left(\sum_{j=1}^{N^{(l-1)}} K_{i,j}^{(l)} * Y_j^{(l-1)} + B_i^{(l)}\right) \tag{2.12}$$

Where $K_{i,j}^{(l)}$ is the applied convolutional kernel, $B_i^{(l)}$ is a bias matrix, h is the activation function illustrated in section 2.1. Recently, there have been several representatives works to improve its effectiveness and efficiency, such as dilated convolution [158], deformable convolution [159], and depth-wise separable convolution [160].

Pooling layer. The pooling layer reduces the data dimensions by sub-sampling the feature maps at one layer by a specific factor and a function. It reduces the number of trainable parameters while it eases the overfitting issue. Pooling operations take a small local region as the input and generate a single value representing this region by defining a downsampling size. The representative value of the receptive field can be computed with a max function ('max-pooling') or an average function ('average pooling'). Another approach to downsampling with a similar effect of pooling is using convolutions with strides [147, 22].

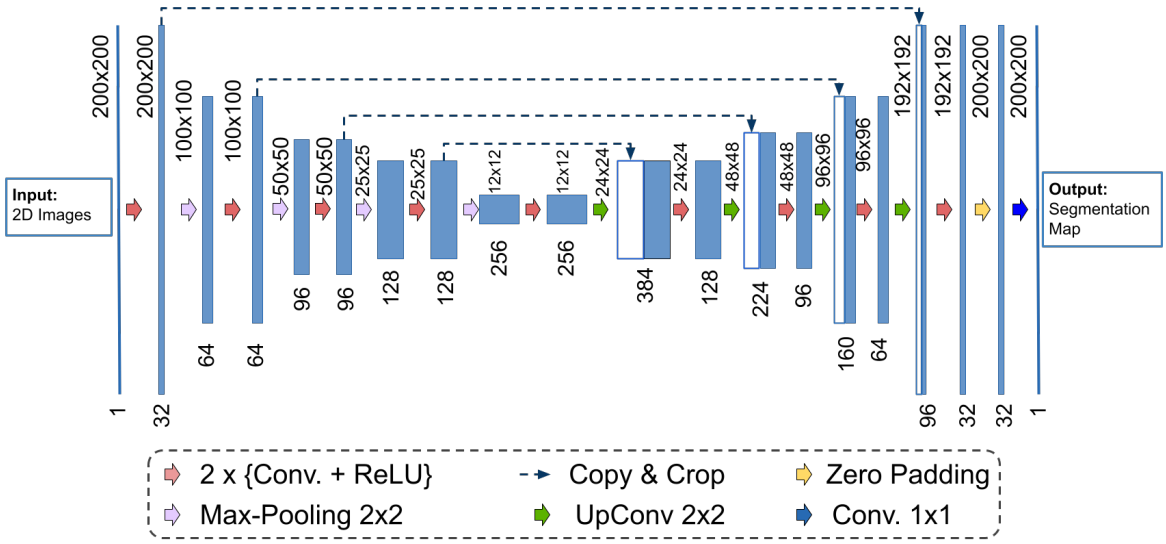


Figure 2.3: A U-shape convolutional neural network architecture for medical image segmentation. Different arrows represent layers and operations. Blue rectangles represent the multi-channel feature maps, and the copied feature maps by skip connect are shown in white rectangles.

Up-sampling layer. The up-sampling layer is commonly used in image segmentation networks to up-sample a feature map to a higher resolution. One approach is *re-sampling and interpolation*. It re-scales an input feature map to the desired size with an interpolation method such as bilinear interpolation. An alternative approach is *unpooling* [161], regarded as the reverse of pooling. In an un-pooling layer, an approximate inverse of the previous pooling layer is obtained by (1) recording the position of each maximum activation value within each pooling region and (2) using this position information prior to reconstructing the previous layer into a higher-resolution feature map. Another method is *transpose convolution* [162]. The transpose convolution is regarded as a reverse of the convolution operation but not a so-called mathematically-defined deconvolution. In the transpose convolution, the kernel is placed over the input, and values of the input are multiplied successively by the kernel weights for producing the up-sampled feature map.

Skip connection. A skip-connection is to connect two layers directly while it skip one or more layers between. It was introduced to relieve the difficulty in non-linearity optimization by propagating a linear component through the neural network layers. The skip-connections propagate the gradient throughout the model, which can alleviate the vanishing gradient problem [151, 152] due to increasing depth of layers. In segmentation networks which are encoder-decoder-like architectures, the skip-connections are utilized to connect each decoder-encoder pair and bring the features with higher spatial resolution from shallow layers of the encoder directly to the layers of the decoder.

Loss function. The loss function measures how far an estimated value is from its reference value and supervises the network training. In image classification tasks, commonly-used loss functions include the categorical cross-entropy loss, Focal loss [51], etc. The categorical cross-entropy loss is defined as:

$$L_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}\delta(y_i = c)\log(P(y_i = c)) \tag{2.13}$$

where N denotes the data number and C represents the categories number. $\delta(y_i = c)$ is the indicator function and $P(y_i = c)$ is the predicted probability by the model.

The Focal loss is a modified cross-entropy loss that weighs the importance of each sample to the loss based on the classification error to tackle the class imbalance issue. It is defined as:

$$L_{FL} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}\delta(y_i = c)\,(1 - P(y_i = c))^{\gamma}\log(P(y_i = c)) \tag{2.14}$$

where γ is a parameter to control the rate of down-weighting easy examples. When $\gamma = 0$, L_{FL} is equivalent to L_{CE}.

Since image segmentation is essentially an image classification task, i.e., it aims to classify each pixel to a label, thus above-mentioned cross-Entropy loss and Focal loss can be used as loss functions for segmentation. Some other loss functions, such as Dice loss, Tversky loss [163], and boundary loss [164] are tailored ones for image segmentation. Assume $p(x_i)$ is the prediction probability of a voxel x_i and $g(x_i)$ is the corresponding reference for the same voxel. The Dice loss is defined as:

$$L_{Dice}(\mathbf{X}) = -\frac{2\sum\limits_{x_i \in \mathbf{X}} p(x_i)g(x_i) + \varepsilon}{\sum\limits_{x_i \in \mathbf{X}} p(x_i) + \sum\limits_{x_i \in \mathbf{X}} g(x_i) + \varepsilon} \tag{2.15}$$

where $\mathbf{X}$ is the training image set, ε is a term to prevent the denominator from being 0.

2.3 Generative adversarial networks

Generative modelling. Generative models take a training set drawn from an unknown distribution p_{data} and return an estimate of that distribution p_{model}. The estimate p_{model} can be evaluated for a particular given value of x. Hence, generative models are able to generate infinite samples from the model distribution p_{model}. Many approaches to generative modeling are based on density estimation which is explicit; given observed training examples of a random variable x, inferring a density function p_{model} that explains the training data, as shown in Fig. 2.5.

In contrast, generative adversarial networks (GANs) [165, 166] are *implicit* generative models that infer the probability distribution $p(x)$ without necessarily representing the density function explicitly. Readers can be more clear about the position of GANs in generative modelling by referring to [167].

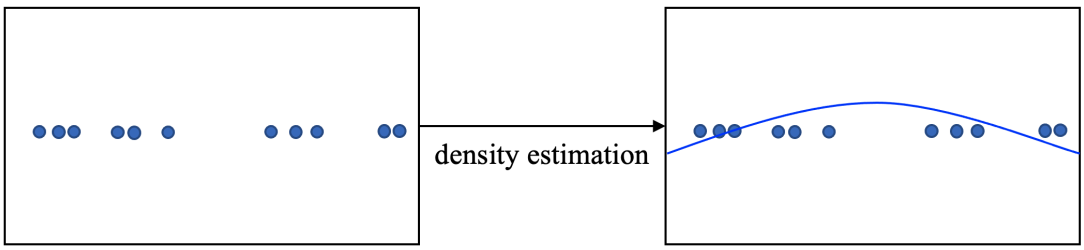


Figure 2.4: An example of generative modelling. Several one-dimensional data points are used to fit a Gaussian density function that explains the observed samples.

Overview of GANs. The basic idea of GANs is to set up a game between two networks. One of them is called *generator*. The generator creates samples that are intended to be from the same distribution as the training data. The other network is called *discriminator* which distinguishes samples to whether they are from the distribution ('real') or not ('fake'). The discriminator learns to classify the inputs into two classes with supervised learning. In the competition, the generator is trained to fool the discriminator. Formally, GANs are a structured probabilistic model containing latent variables z and observed variables x. The two nets in the game are represented by two differentiable functions. The discriminator is a function D that takes x as

input and uses $\theta^{(D)}$ as parameters. The generator is defined by a function G that takes z as input and uses $\theta^{(G)}$ as parameters.

Both nets are optimized with cost functions. The discriminator aims to minimize an objective function that involves two nets $J^{(D)}(\theta^{(D)}, \theta^{(G)})$, and must do so while controlling only $\theta^{(D)}$. The generator wishes to minimize $J^{(G)}(\theta^{(D)}, \theta^{(G)})$ with respect to only $\theta^{(D)}$. This is because each network's cost depends on the other network's parameters, but the two networks are independent in parameters.

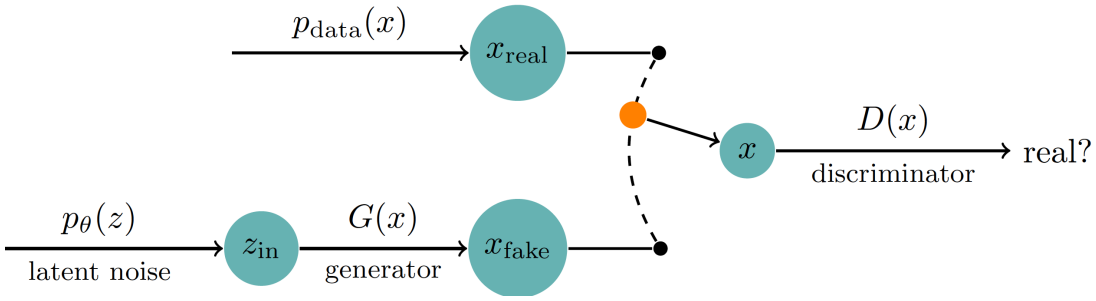


Figure 2.5: A schematic view of GANs as an implicit generative model.

2.4 Transfer learning and domain adaptation

Transfer learning solves a basic problem of insufficient training data in a given task. It transfers the knowledge from the source domain to a target domain by relaxing the assumption that the training data and the test data must be independent and identically distributed. For example, transferring knowledge from natural image domain to medical domain [168] can be highly helpful due to insufficient training data in the latter domain. The procedure of transfer learning is illustrated in Figure 2.6.

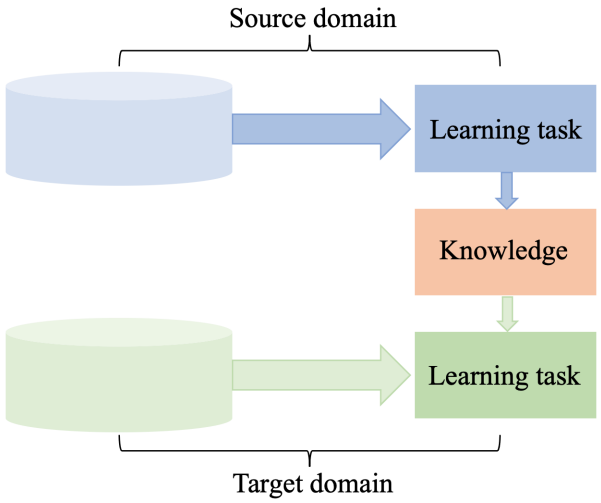


Figure 2.6: The procedure of transfer learning between source and target domains.

A domain can be represented by $\mathscr{D} = \{\chi, P(X)\}$, which consists of two parts: an image space χ and an edge probability distribution $P(X)$ where $X = \{x_1, ..., x_n\} \in \chi$. A task can be defined by $T = \{y, f(x)\}$. It contains two parts: a label space y and a target prediction function $f(x)$. $f(x)$ can also be regarded as a conditional probability function $P(y|x)$. Given a source domain and a target domain with data X and its distribution $P(X)$, $\mathcal{D}_s$=$\{X_s, P(X_s)\}$, $\mathcal{D}_t$=$\{X_t, P(X_t)\}$ and a shared *output* space $\mathcal{Y} = \{Y\}$, transfer learning can be formally defined as follows:

Definition of transfer learning: Given a target learning task $\mathbf{T_t}$ based on data from target domain $\mathbf{D_t}$, and we can get the help from $\mathbf{D_s}$ for the source learning task $\mathbf{T_s}$. Transfer learning aims to improve the performance of predictive function $\mathbf{f_T}$ for learning task $\mathbf{T_t}$ by discovering and transferring latent knowledge from $\mathbf{D_s}$ and $\mathbf{T_s}$, where $D_s = D_t$ or $T_s = T_t$. In addition, in the most case, the size of D_s is much larger than the size of D_t, N_s is much larger than N_t.

Domain adaptation is a sub-field of transfer learning. In domain adaptation, it is assumed that the source and target domains share the same feature space but in different distributions; in contrast, transfer learning includes cases where the target domain's feature space is different from the source feature space or spaces. A predictive model $f(\cdot)$ which approximates $P(Y|X)$ trained on the source domain $\mathcal{D}_s$ is likely to degrade on the target domain $\mathcal{D}_t$ which presents a domain shift. Among existing works, one of the key ideas is to match the input space for both domains in the feature space so that the mapping can be invariant to the inputs [169].

3 Effective brain lesion and claustrum segmentation with ensembles of deep nets

This chapter has been published as two **peer-reviewed journal publications**:

[1] **H. Li**, G. Jiang, J. Zhang, R. Wang, Z. Wang, W.-S. Zheng, and B. Menze. "Fully convolutional network ensembles for white matter hyperintensities segmentation in MR images". In: ***NeuroImage*** 183 (2018), pp. 650–665

[2] **H. Li**, A. Menegaux, B. Schmitz-Koep, A. Neubauer, F. J. Bäuerlein, S. Shit, C. Sorg, B. Menze, and D. Hedderich. "Automated claustrum segmentation in human brain MRI using deep learning". In: ***Human Brain Mapping*** 42.18 (2021), pp. 5862–5872

Synopsis: The above two works develop effective deep learning based methods with ensemble learning for brain structure segmentation. Specifically, publication ♯1 establishes a state-of-the-art algorithm for white matter hyperintensities segmentation from FLAIR and T1 scans, which won a segmentation grand challenge at MICCAI 2017. The ensemble of single deep neural nets significantly boosts segmentation performance on multi-center MRI scans, especially improves the segmentation of small lesions. Technical details, methodological choice and analysis are presented. Publication ♯2 further develops the methodology by incorporating multi-view information and improve the segmentation accuracy of human claustrum. Both two segmentation methods reach expert-level performance, are evaluated in a cross-center manner and have potential in real-world clinical practice.

Contributions of thesis author: algorithm design and implementation, computational experiments and composition of manuscript.

NeuroImage 183 (2018) 650–665

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Fully convolutional network ensembles for white matter hyperintensities segmentation in MR images

Hongwei Li [a,b,c], Gongfa Jiang [a], Jianguo Zhang [b,*,1], Ruixuan Wang [a,**,2], Zhaolei Wang [a], Wei-Shi Zheng [a], Bjoern Menze [c]

[a] School of Data and Computer Science, Sun Yat-sen University, China
[b] Computing, School of Science and Engineering, University of Dundee, UK
[c] Department of Computer Science, Technical University of Munich, Germany

ARTICLE INFO

Keywords:
White matter hyperintensities
Brain lesion segmentation
MICCAI WMH segmentation challenge
Deep learning
Ensemble models

ABSTRACT

White matter hyperintensities (WMH) are commonly found in the brains of healthy elderly individuals and have been associated with various neurological and geriatric disorders. In this paper, we present a study using deep fully convolutional network and ensemble models to automatically detect such WMH using fluid attenuation inversion recovery (FLAIR) and T1 magnetic resonance (MR) scans. The algorithm was evaluated and ranked 1st in the WMH Segmentation Challenge at MICCAI 2017. In the evaluation stage, the implementation of the algorithm was submitted to the challenge organizers, who then independently tested it on a hidden set of 110 cases from 5 scanners. Averaged dice score, precision and robust Hausdorff distance obtained on held-out test datasets were 80%, 84% and 6.30 mm respectively. These were the highest achieved in the challenge, suggesting the proposed method is the state-of-the-art. Detailed descriptions and quantitative analysis on key components of the system were provided. Furthermore, a study of cross-scanner evaluation is presented to discuss how the combination of modalities affect the generalization capability of the system. The adaptability of the system to different scanners and protocols is also investigated. A quantitative study is further presented to show the effect of ensemble size and the effectiveness of the ensemble model. Additionally, software and models of our method are made publicly available. The effectiveness and generalization capability of the proposed system show its potential for real-world clinical practice.

1. Introduction

Small vessel diseases are mainly systemic disorders that affect various tissues and organs of human body. These diseases are thought to be the most frequent pathological neurological process and have a crucial role in at least three fields: stroke, dementia and aging (Pantoni, 2010).

White matter lesions characterized by bilateral, mostly symmetrical hyperintensities, are commonly seen on FLAIR MRI of clinically healthy elderly people; furthermore, they have been repeatedly associated with various neurological and geriatric disorders such as mood problems and cognitive decline (Kim et al., 2008; Debette and Markus, 2010). Manual delineation of WMH area, as shown in Fig. 1, is a reliable way to assess white matter abnormalities but this process is laborious and time-consuming for neuroradiologists and shows high intra-rater and inter-rater variability (Grimaud et al., 1996).

Computer vision and machine learning techniques have increasingly shown a promising road for automatic diagnosis of diseases through medical imaging. By analyzing imaging data in a statistical manner, many image processing algorithms dealing with brain lesions generalize well within closely related applications, for example, in the segmentation of WMH, multiple sclerosis (MS), tumors, stroke, and even traumatic brain injury. Although various computer-aided diagnosis systems have been proposed for these different brain lesion segmentation tasks, the reported results are largely incomparable due to different datasets and evaluation protocols.

Van Leemput et al. (2001) presented an early attempt at developing

* Corresponding author.
** Corresponding author.
E-mail addresses: j.n.zhang@dundee.ac.uk (J. Zhang), wangruix5@mail.sysu.edu.cn (R. Wang).
[1] * indicate the first corresponding author.
[2] ** indicate the second corresponding author.

https://doi.org/10.1016/j.neuroimage.2018.07.005
Received 25 December 2017; Received in revised form 30 June 2018; Accepted 2 July 2018
Available online 18 August 2018

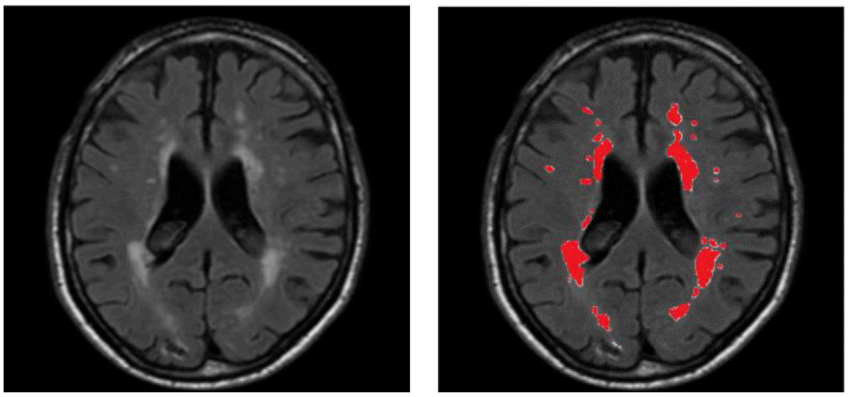


Fig. 1. A sample of MR slice from FLAIR modality (left), and its corresponding manual annotation of WMH by a neuroradiologist (right).

an unsupervised-learning-based segmentation system to detect multiple sclerosis lesions from large datasets of T1-weighted (T1), proton density-weighted (PD) and T2-weighted (T2) scans. The method simultaneously estimates the parameters of a stochastic model for normal brain MR images and detects MS lesions as outliers of the model. Anbeek et al. (2004) developed a supervised-learning-based automated system using T1, inversion recovery, PD, T2 and fluid attenuation inversion recovery (FLAIR) scans. Intensity and 3D spatial features were extracted from the voxels and are used to train a k-nearest neighbors classifier. Dyrby et al. (2008) used artificial neural networks based on intensity and spatial information, in which six optimized networks were produced to investigate the impact of different input modalities on WMH segmentation. Beare et al. (2009) developed a method that searched for WMHs per-region instead of per-voxel. The region-based features are combined with an adaptive boosting statistical classifier. Geremia et al. (2010, 2011) were the first to address the MS lesion segmentation in a straightforward learning approach using context-rich, symmetry and local spacial features and random forest. Simões et al. (2013) built the intensity histogram of FLAIR by a Gaussian mixture model. Then the probability of a voxel depends on not only the voxel's intensity but also on its neighbors' current class probabilities. Schmidt et al. (2013) contributed an open source tool for the segmentation of hyperintensities that integrates with the popular SPM package. Yoo et al. (2014) developed an intensity-based, monospectral segmentation method in which the optimal intensity threshold on FLAIR images varied with WMH volume. Very recently, Ghafoorian et al. (2017) integrated the anatomical location information into the convolutional neural networks (CNN), in which several deep CNN architectures that consider multi-scale patches or take explicit location features were proposed. Moeskops et al. (2017) proposed a patch-based deep CNN to segment brain tissues and WMH in MR images.

In computing research, benchmarking on specific problems is an effective way to fairly compare state-of-the-art methods. There have been several related benchmarks on automated segmentation of different brain tissues in MR images in the field of medical image analysis. The Multiple Sclerosis Lesion Segmentation Challenge 2008 organized by Styner et al. (2008) is one of the early contests for comparing the methods for automatic extraction of MS lesions from T1, T2 and FLAIR MRI data. The Ischemic Stroke Lesion Segmentation Challenge (ISLES) from 2015 to 2017 organized by Maier et al. (2017) provides a platform for fair comparison of stroke lesion segmentation algorithms. The Multi-modal Brain Tumor Segmentation Challenge (BRATS) organized by Menze et al. (2015) draws much attention since 2012 which focuses on segmentation of low- and high-grade gliomas, more recently, prediction of patient overall survival. Different from the above tasks, WMH tend to have consistent patterns such as significant symmetry, but they are more scattered, often with some regions of very small size and irregular shapes. Furthermore, compared with other brain tissue segmentations, WMH segmentations are more likely to be susceptible to the presence of motion artefacts and other brain abnormalities, such as brain infarcts (Gouw et al., 2010).

The *WMH Segmentation Challenge 2017*[3] was held to compare state-of-the-art algorithms in conjunction with the 20th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI, 2017). This paper describes our winning entry to this challenge in detail, which was evaluated by the organizers on clinical datasets. The algorithm was containerized and applied to the test datasets by the challenge organizers, while the test sets remained unseen to us and other contestants. The test set includes 110 secret cases from five different MR scanners world-widely from three hospitals in the Netherlands and Singapore. Our approach to detecting WMH in MR images is based on an ensemble of convolution-deconvolution architecture (Long et al., 2015) with long-range connections (Ronneberger et al., 2015) which simultaneously classifies each pixel and locates objects of an input image. In our system (as shown in Figs. 2, 4, 5), we implement a network architecture with 19 layers that are optimized for classifying and localizing the WMH. Ensemble models trained with random parameter initializations and shuffled data are employed for voting the pixel labels in the final evaluation.

This paper is organized as follows. Section 2 describes the datasets, rating criteria, five evaluation metrics on segmentation performance and rank method of the challenge. Section 3 presents in detail each component of our method and how some key parameters are optimized. Section 4 evaluates the proposed system on the public training dataset (60 cases) and reports results for the hidden held-out dataset (110 cases). Section 5 discusses different aspects of our winning method. This includes the motivation to use 2D model instead of 3D one, a novel *cross-scanner* study on how the combination of modalities and data augmentation strengthen the generalization capability to unseen scanners. Furthermore, evaluation on the adaptability to various scanners as well as quantitative analysis on the optimal number of ensemble models are performed.

2. Materials

This section mainly describes the WMH Segmentation Challenge, datasets, evaluation metrics and rank method which are referred to in the rest of the article.

2.1. MICCAI WMH segmentation challenge overview

The challenge organized as a joint effort of the *UMC Utrecht, VU Amsterdam* and *NUHS Singapore*, aims at, for the first time, benchmarking methods for automatic WMH segmentation of presumed vascular origin. Sixty cases from three centers were released as a public training set for participants to build and evaluate their algorithms. One hundred and ten

[3] http://wmh.isi.uu.nl/.

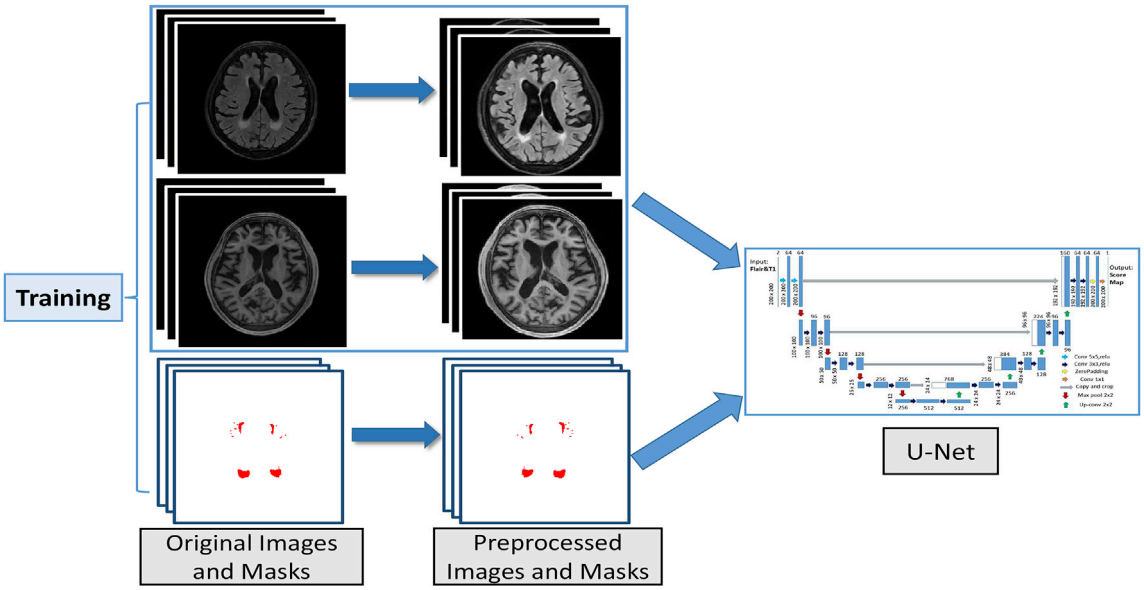


Fig. 2. Overall framework of the training stage.

Table 1
Characteristics of *MICCAI WMH Challenge* dataset. The training set consists 60 subjects' data from 3 scanners and the test set includes 110 cases from 5 scanners (two of them are not represented in the training set).

Datasets	Scanners Name	Voxel Size (m^3)	Size of FLAIR Scans	Train	Test
Utrecht	3T Philips Achieva	0.96 × 0.95 × 3.00	240 × 240 × 48	20	**30**
Singapore	3T Siemens TrioTim	1.00 × 1.00 × 3.00	252 × 232 × 48	20	**30**
GE3T	3T GE Signa HDxt	0.98 × 0.98 × 1.20	132 × 256 × 83	20	**30**
GE1.5T	3T Philips Ingenuity	1.04 × 1.04 × 0.56	secret	–	**10**
PETMR	1.5T GE Signa HDxt	1.21 × 1.21 × 1.30	secret	–	**10**

hidden cases from five scanners are used by the organizers to test the algorithms. Notably, all algorithms are containerized by *Docker* (Merkel, 2014) to guarantee that the test data remains secret and cannot be included in any way in the training procedure of the techniques. Twenty international teams participated, and further information including training data and the results on test set are made public via the following url: http://wmh.isi.uu.nl/results/.

2.2. Datasets

In all reported experiments, we relied on the publicly available dataset from the MICCAI WMH Challenge. Properties of the data are summarised in Table 1. A notable feature is that the images were acquired from five different scanners from three hospitals in the Netherlands and Singapore. As shown in Table 1, there exists large difference in acquisition settings; in particular voxel sizes of the captured images differ significantly among the five scanners. For each subject, a 3*D* T1-weighted image, and a 2*D* multi-slice FLAIR image were provided. Since the manual reference standard is defined on the FLAIR image, a 2*D* multi-slice version of the T1 image was generated by re-sampling the 3*D* T1-weighted image to match with the FLAIR one. Finally, the pre-processed images were corrected for bias field inhomogeneities using *SPM12*.[4] The 3D FLAIR image was resampled to a slice-thickness of 3.00 mm and there is no gap between slices.

The dataset consists of in total 170 subjects with FLAIR and T1 MR images from five different scanners along with their binary masks. The images from 60 subjects were made available during the training stage. The images from the remaining 110 subjects were used as the hidden test set to evaluate performance of methods submitted to the challenge. Notably, the test set also includes images of 20 subjects captured by other two *unseen* scanners, which were not used to capture images for training. This dataset setting encourages the participants to submit algorithms that could be robust to unseen scanners.

2.3. Evaluation metrics and rank method

Five different metrics are used by the challenge organizers to compare and rank the methods by different teams; those metrics evaluate the segmentation performance in different aspects.

Given a ground-truth segmentation map G and a segmentation map P generated by an algorithm, the five evaluation metrics are defined as follows.

2.3.1. Dice similarity coefficient (DSC)

$$\mathrm{DSC} = \frac{2(G \cap P)}{|G| + |P|} \tag{1}$$

This measures the overlap in percentage between G and P.

2.3.2. Hausdorff distance (95th percentile)

Hausdorff distance is defined as:

$$H(G,P) = max\left\{ \sup_{x \in G} \inf_{y \in P} d(x,y), \sup_{y \in P} \inf_{x \in G} d(x,y) \right\} \tag{2}$$

where $d(x, y)$ denotes the distance of x and y, *sup* denotes the supremum and *inf* for the infimum. This measures how far two subsets of a metric space are from each other. As used in this challenge, it is modified to obtain a robustified version by using the 95th percentile instead of the

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

maximum (100th percentile) distance.

2.3.3. Average volume difference (in percentage)

Let V_G and V_P be the volume of lesion regions in G and P respectively. Then the Average Volume Difference (AVD) in percentage is defined as:

$$\mathrm{AVD} = \frac{|V_G - V_P|}{V_G} \tag{3}$$

2.3.4. Sensitivity for individual lesions (recall)

Let N_G be the number of individual lesions delineated in G, and N_P be the number of correctly detected lesions after comparing P to G. Each individual lesion is defined as a 3D connected component. Then the recall for individual lesions is defined as:

$$\mathrm{Recall} = \frac{N_P}{N_G} \tag{4}$$

2.3.5. F1-score for individual lesions

Let N_P be the number of correctly detected lesions after comparing P to G. N_F be the number of wrongly detected lesions in P. Each individual lesion is defined as a 3D connected component. Then the F1-score for individual lesions is defined as:

$$\mathrm{F1} = \frac{N_P}{N_P + N_F} \tag{5}$$

The full source code for computing the evaluation metrics can be found on: https://github.com/hjkuijf/wmhchallenge/blob/master/evaluation.py.

For each team, the values of those five metrics were computed by the organizers independently. For each evaluation metric, the performances of all of the teams were sorted from best to worst. Then a calibrated score for each team was computed by normalising its performance w. r.t the range of all the actual performances for that metric. Thus the best team was assigned a rank score of one, while the worst team got a rank score of zero. Other teams received a score of between (0,1). Finally, for each team, the rank scores of the five metric were averaged into the final score, being the overall performance of that team. For consistency, when presenting the results of the challenge, we follow exactly the same ranking criteria.

3. Methods

3.1. Further preprocessing

A further preprocessing on top of the basic preprocessing steps pursued by the organizers (Section 2.2) plays an important role in our overall framework. We aim at employing a simple and effective preprocessing step on both training and held-out testing set. It is motivated by three objectives: 1) to guarantee a uniform size of all data for deep convolutional networks in the training and test stage, 2) to normalize voxel intensity to reduce variation across subjects. and 3) to equip the CAD system with desired invariance and robustness. We enforce these desired data properties by implementing further steps in the training of our algorithm: 1) cropping or padding each axial slice to a uniform size, 2) Gaussian normalization on the brain voxel intensity, and 3) data augmentation on the processed images. Most of these steps are performed for both FLAIR and T1 modalities and for both the training and test stages. Data augmentation was performed only during the training stage.

Firstly, all the axial slices were automatically cropped or padded to 200×200, in order to guarantee a uniform size for input to the deep-learning model. Secondly, Gaussian normalization was employed to normalize the intensity distributions for each 3D scan. This includes three steps. Firstly, a threshold was empirically set to obtain an initial binary brain mask. Secondly, for each axial slice of the obtained binary masks, the largest connected component was selected. Thirdly, the holes inside the connected component was filled using morphology operations. Thus a final brain mask was obtained for each slice. For each 3D scan, Gaussian normalization was then employed to rescale the voxel intensities *within* each individual's brain mask.

The thresholds for creating the brain masks were empirically set to 70 for FLAIR and 30 for T1 respectively. It was noted that several methods submitted for the contest extracted the brain using common tools such as BET (Smith, 2002), where the skull was also removed. However, we found the removal of skull has little effect on the performance of the proposed system.

3.1.1. Data augmentation

Data augmentation is an effective way to equip the deep networks with desired invariance and robustness properties when training data are limited. In case of MR images among different subjects and scanners, due to variations of head orientations, voxel sizes and WMH distribution, we primarily need rotation and scale invariance as well as robustness to shear transformation. For each axial slice, three transformations including rotation, shear mapping and scaling were applied, each within a parameter range. The parameter range represents the variation in different aspects between subjects in clinical practice; for example, rotation of brain is in the range of [-15°, 15°]. Table 2 lists the parameter range for each of the three transformations. It should be noted that the scaling used in the training of the algorithm was in the range of (0.9, 1.1), representing the range of voxel size ratios in the training data sets (Table 1), while some test sets had noticeable larger ratios (a factor of 1.21 between the PETMR and the Singapore data set). This indicates the robustness of our approach, but also leaves potential room for improvement in future studies exploring the optimal scaling of the data during training.

Fig. 3 shows an example of the resulting slices after applying the transformations. After data augmentation, we obtain a dataset four times larger than the original one.

3.2. Fully convolutional network

3.2.1. 2-D convolutional network architecture

Convolutional neural network has proven to be an effective computational model for automatically extracting image features. Recently the fully convolutional networks (FCN) (Long et al., 2015) and their its extensions (Milletari et al., 2016) have been used for medical images segmentation. We build a variant of FCN architecture based on U-Net (Ronneberger et al., 2015), which takes as input the axial slices of two modalities from the brain MR scans during both training and testing. Our network is shown in Fig. 4. For each patient, the FLAIR and T1 modalities are fed into the U-Net jointly as a two-channel input. It consists of a down-convolutional part that shrinks the spatial dimensions (left side), and up-convolutional part that expands the score maps (right side). The skip connections between down-convolutional and up-convolutional were employed.

In this model, two convolutional layers are repeatedly employed, each followed by a rectified linear unit (ReLU) and a 2×2 max pooling operation with stride 2 for downsampling. At the final layer a 1×1 convolution is used to map each 64-component feature vector to two classes. In total the network contains 19 convolutional layers. Convolutional layers with 3×3 kernel size are heavily used in our model. Different from the basic architecture of the recent work (Ronneberger et al., 2015), for the first two convolutional layers, kernel size 3×3 is

Table 2
Parameters range used for data augmentation. The value range in column *Shearing* indicates the shear angle. The value range in column *scaling* indicates the scale factor.

Methods	Rotation	Shearing	Scaling (x & y)
Parameters	[-15°, 15°]	[-18°, 18°]	[0.9, 1.1]

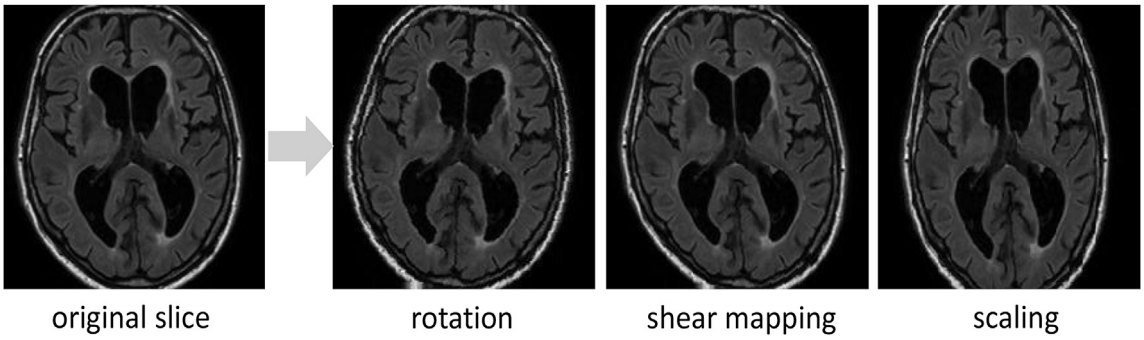


Fig. 3. An example of data augmentation result. From left to right: the original axial slice, slice after rotation, slice after shear mapping and slice after scaling.

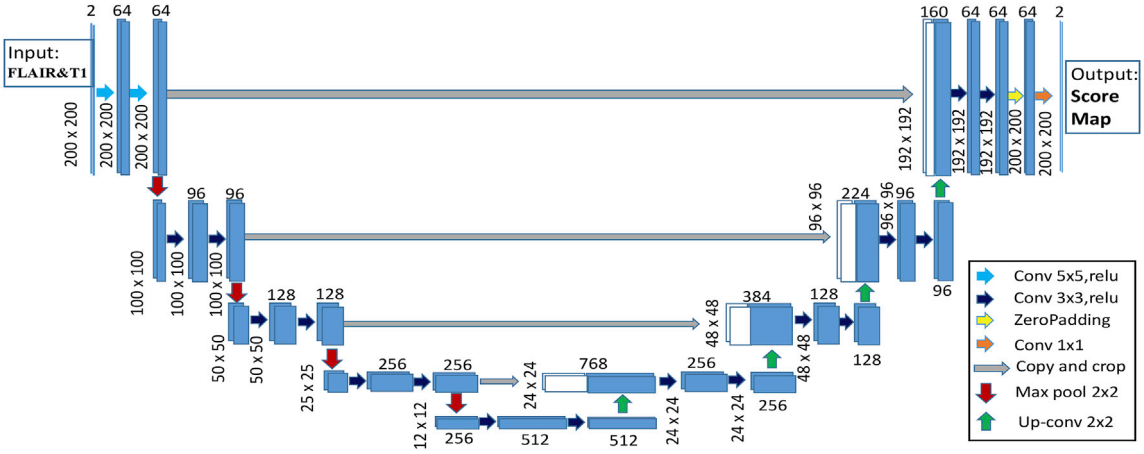


Fig. 4. 2D Convolutional Network Architecture. It consists of a shrinking part (left side) and an expansive part (right side) to detect and locate*WMH* respectively. The input includes FLAIR and T1 channel.

replaced with size 5×5 in order to handle different transformations. This is motivated by a recent study (Peng et al., 2017) suggesting that large kernel size should be adopted in the network architecture. This step could enable dense connections between feature maps and per-pixel classifiers, enhancing the capability of a network to handle different transformations.

3.2.2. *Dice loss*

In the task of WMH segmentation, the numbers of positives and negatives are highly unbalanced. One of the solutions to tackle this issue is to use Dice loss (Milletari et al., 2016) as the loss function for training the model. The formulation is as follows.

Let $G = \{g_1, \ldots, g_N\}$ be the ground-truth segmentation probabilistic maps (gold standard) over N slices, and $P = \{p_1, \ldots, p_N\}$ be the predicted probabilistic maps over N slices. The Dice loss function can be expressed as:

$$DL = -\frac{2\sum_{n=1}^{N}|p_n \circ g_n| + s}{\sum_{n=1}^{N}(|p_n| + |g_n|) + s} \tag{6}$$

where $\circ$ represents the entrywise product of two matrices, and $|\cdot|$ represents the sum of the entries of matrix. The s term is used here to ensure the loss function stability by avoiding the division by 0, i.e., in a case where the entries of G and P are all zeros. s was set to 1 in our experiments.

3.3. *Ensemble FCNs*

Ensemble techniques are helpful to reduce over-fitting problems of a complex model on the training data (Opitz and Maclin, 1999). It combines multiple learning models to obtain better predictive performance than any of the constituent learning algorithms alone. There exists various work using ensembles of deep learning models in computer vision and medical image analysis. Krizhevsky et al. (2012) and Simonyan and Zisserman (2014) achieved top performance in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012 and 2014 by averaging multiple deep CNNs with same architectures. He et al. (2016) won the first place with an ensemble of six Residual Networks with different depths in ILSVRC (2015). Kamnitsas et al. (2017a) won the brain tumor segmentation challenge (BraTs) 2016 by aggregating different segmentation networks. In this work, we propose to address the automated WMH segmentation problem by an ensemble approach to combine several models with same architecture in a carefully designed pipeline. We further show the effectiveness of the ensemble model via a quantitative analysis in Sections 5.6 and 5.7.

The intention to use ensemble models includes two aspects: 1) different models could learn different attributes of the training data during the batch learning processing, thus the ensemble of them could boost the segmentation results; 2) bias-variance trade-off. Assume that network model error is due to bias and variance. If the variance of model decrease, then the overall error would likely decrease. Here we aimed to lower the variance by averaging the model outputs. A FCN with millions of parameters, over-trained on different bootstrapped/subsampled training sets would qualify for unbiased and highly variant models. We further discussed in Section 5.6 that ensemble model served as the typical bias-variance trade-off.

As shown in Fig. 5, n U-Net models with same architecture are trained with random parameter initialization and shuffled data in the batch learning. For each of the n U-Net models, when given a test image, a

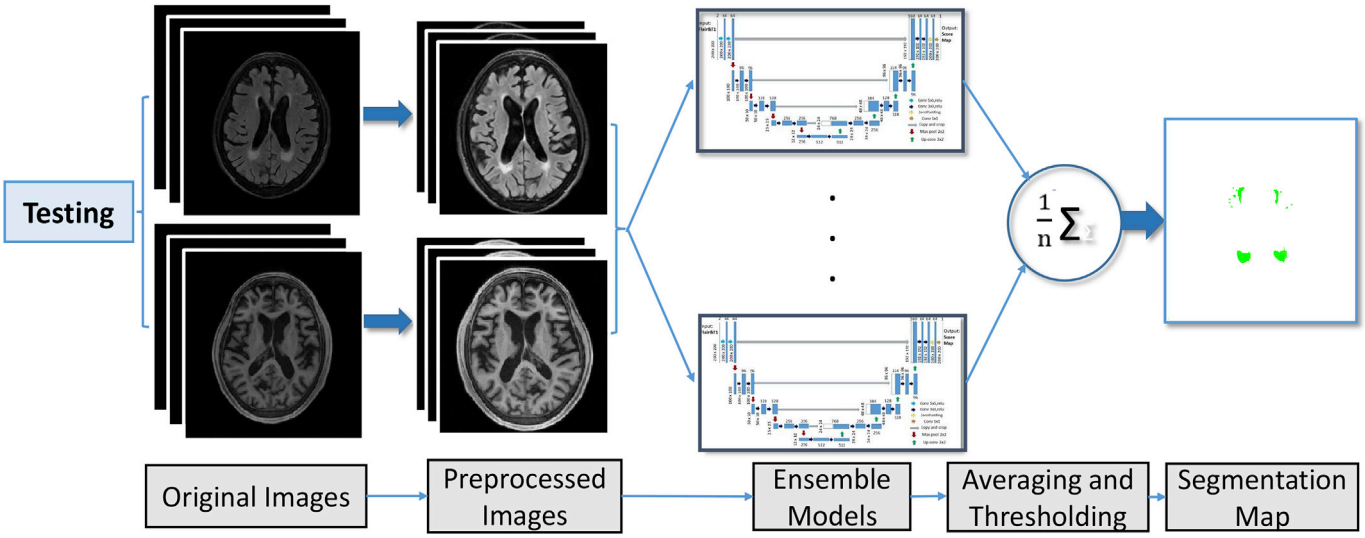


Fig. 5. Overall framework for the testing stage.

probability segmentation map will be generated by that model. Then the resulting n maps will be averaged. Finally an empirically-picked threshold will be used to transform the scores map into a binary segmentation map.

3.4. Post-processing

The post-processing includes two aspects: 1) cropping or padding the segmentation maps with respect to the original size, i.e., an inverse operation to the step described in Section 3.1; 2) removing some anatomically unreasonable artefact in the axial slices. For the purpose of removing unreasonable detections (e.g., WMH will not appear in the first few axial slices containing neck and last few axial slices containing skull), we employed a simple strategy: if there exists detected WMH in the first m slices and last n ones of a brain along the z-direction, then the WMH regions were considered as false positive and would be removed. Empirically, m and n were set to 10% of the number of slices for each scan. The codes and models of the proposed system is made publicly available in *GitHub*.[5]

4. Results

In this section we report the segmentation performances on both the public training dataset and the held-out test set and compare to other teams' methods presented during the challenge. Detailed segmentation results of the 20 teams on the 110 secret cases are available in the following url: http://wmh.isi.uu.nl/results/.

For reported results, the binary segmentation maps were evaluated using the five metrics described in Section 2: dice similarity coefficient, Hausdorff distance (95p), averaged volume difference, lesions recall and lesions F1-score. The U-Net hyper parameters were set as follows: batch size for computing the training loss was set to 30; learning rate was set to 0.0002; the number of epochs was set to 50. The number of models in the ensemble was set to 3. Section 5.2 further evaluates and analyses the effects of some key parameters on segmentation performance.

4.1. Results on held-out test dataset

The proposed system was announced to be the winning method of the challenge after being independently tested on 110 hidden cases from 5 scanners by the organizers. The overall ranking was based on the average of the rank scores computed for each metric. For the testing stage, deep fully convolutional networks were learned on the whole public training dataset consisting of 60 cases. Table 3 shows the segmentation performance of our submitted system on the held-out test set with its 5 subsets, each containing cases from the different scanners and sites. Table 4 compares our method to other top performing teams. Notably, the top-5 methods all used deep learning techniques, briefly described in Table 5. The proposed FCN ensemble achieved, on average, the highest dice similarity coefficient, smallest Hausdorff distance and best lesion recall. For the 20 cases from unseen scanners *AMS GE1*.5T and *AMS PETMR*, our method achieved the highest lesion recalls of 90% and 84% respectively. We will discuss in Section 5 how each key component of our method, especially the model ensemble, contributes to the improvement on the generalization capability.

4.2. Leave-one-subject-out evaluation on public training dataset

To test the generalization performance of our system across different subjects, we conducted an experiment on the public training datasets (60 subjects) in a leave-one-subject-out setting. Specifically, we used the subject IDs to split the public training dataset into training and validation sets. There were 60 different subjects available. In each split, we used slices from 59 subjects for training, and the slices from the remaining subject for testing. This procedure was repeated until all of the subjects are used as testing.

Fig. 6 plotted the distributions of segmentation performances on scans from the three scanners, with each sub-figure showing performances using one of the five metrics. It could be observed that the segmentation performance on *Utrecht* was relatively poor. A few outliers (hard examples) were found in *Utrecht* which appeared to contain

Table 3
Results of our method on the heldout sets from the five different scanners. ↓ indicates that smaller value represents better performance. The last row shows the rank scores of our method w.r.t the 20 teams for each of the five metrics, with 0 = *best*, and 1 = *worst*.

Scanners	DSC	H95 ↓	AVD ↓	Recall	F1
Utrecht (n = 30)	0.80	7.22	18.35	0.81	0.72
Singapore (n = 30)	0.83	4.50	19.95	0.85	0.78
GE3T (n = 30)	0.79	4.04	24.46	0.83	0.79
AMS GE1.5T *(n = 10)*	0.77	10.24	36.86	0.90	0.80
AMS PETMR (n = 10)	0.72	11.84	15.54	0.84	0.65
weighted average	**0.80**	**6.30**	21.88	**0.84**	0.76
rank scores [0–1]	0.000	0.000	0.004	0.000	0.034

[5] https://github.com/hongweilibran/wmh_ibbmTum.

Table 4
Performance of top-5 methods among the 20 teams. The cells in gray shading indicate the best segmentation performance on each metric. The overall ranking is based on the average of the rank scores on each metric as shown in last row of Table 3. ↓ indicates that smaller value represents better performance.

Teams	Rank/score	DSC	H95↓	AVD↓	Recall	F1
Ours	**1**/0.038	0.80	6.30	21.88	0.84	0.76
cian	2/0.181	0.78	6.82	21.72	0.83	0.70
nlp_logix	3/0.243	0.77	7.16	18.37	0.73	0.78
nih_cidi_2	4/0.302	0.76	7.02	27.98	0.81	0.70
nic – vicorob	5/0.369	0.77	8.28	28.54	0.75	0.71

Table 5
Brief description of top-five methods.

Team Names	Brief Description of Methods
sysu_media(ours)	Fully convolutional network ensembles.
cian	Multi-dimensional gated recurrent units based on recurrent neural networks.
Nlplogix	Two densely connected deep convolutional neural networks.
nih_cidi_2	Traditional deep fully convolutional neural network and graph refinement.
nic – vicorob	A cascade of three convolutional neural networks.

relatively more small lesions and blurred slices after checking the original slices and segmentation results. Section 5 presents a further analysis of these outliers, revealing the challenge of WMH segmentation task. In general, the averaged dice similarity coefficient, Hausdorff distance and lesion recall achieved by the proposed system on 60 cases were 87%, 3.6 mm and 85%, respectively. This shows its effectiveness in aspects of overlapping, localization accuracy and overall lesion detection. Table S1 in the supplemental material reports extensive results allowing comparison on every case of the public training dataset.

4.3. Cross-scanner evaluation

To further evaluate the generalization performance to unseen scanners, firstly we presented a study of cross-scanner analysis on public training set containing 60 cases from three scanners. Then we directly re-ranked and compared the cross-scanner segmentation performance of all teams' methods on the two unseen scanners.

For the cross-scanner analysis, we used the scanner IDs to split the 60 cases into training and test sets. In each split, the slices of 40 subjects from two scanners were used as training set while the slices of 20 subjects from the remaining scanner were used for validation set. This procedure was repeated until all the scanners are used as validation set. For comparing the cross-scanner performance with other state-of-the-art methods, we calculated averaged performances of all teams on the two unseen scanners *AMS GE1*.5T and *AMS PETMR*. Then each team's ranking score was calculated using the same rank method introduced in Section 2.3.

Fig. 7 plots the distributions of segmentation performances on cases from each scanner being tested in turn, with each sub-figure showing performances using one of the five metrics. In general, for every 20 cases from each of the three testing scanners in the cross-scanner evaluation, the segmentation result between each other was comparable, showing our system is robust to unseen scanners. It could be observed that the segmentation performance on dataset *GE3T* was relatively poor. This could be explained that the voxel size of cases in *GE3T* has a significant difference from that captured by two other scanners. Combination of modalities will be discussed in Section 5.3 Table 6 compares the segmentation performances of the top performing teams on two unseen scanners. Our method achieved, on average, the best Dice similarity coefficient and lesion recall of 74.5% and 87% respectively and runner-ups on other three metrics.

5. Discussion

In this section, we further present relevant results obtained on the training data and that impacted on our design choices.

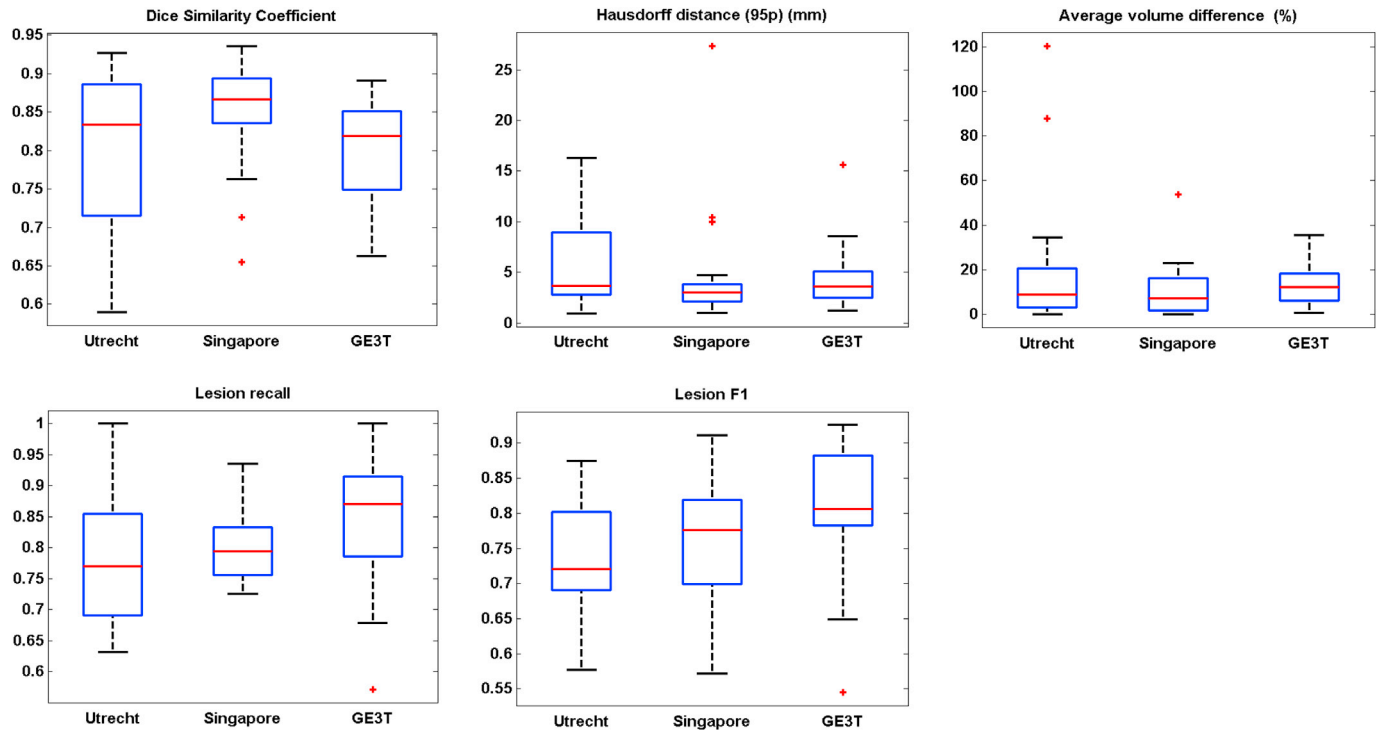


Fig. 6. Box plots of leave-one-subject-out evaluation on the public training data. Each box plot summarizes the segmentation performance on images from one scanner using one specific metric.

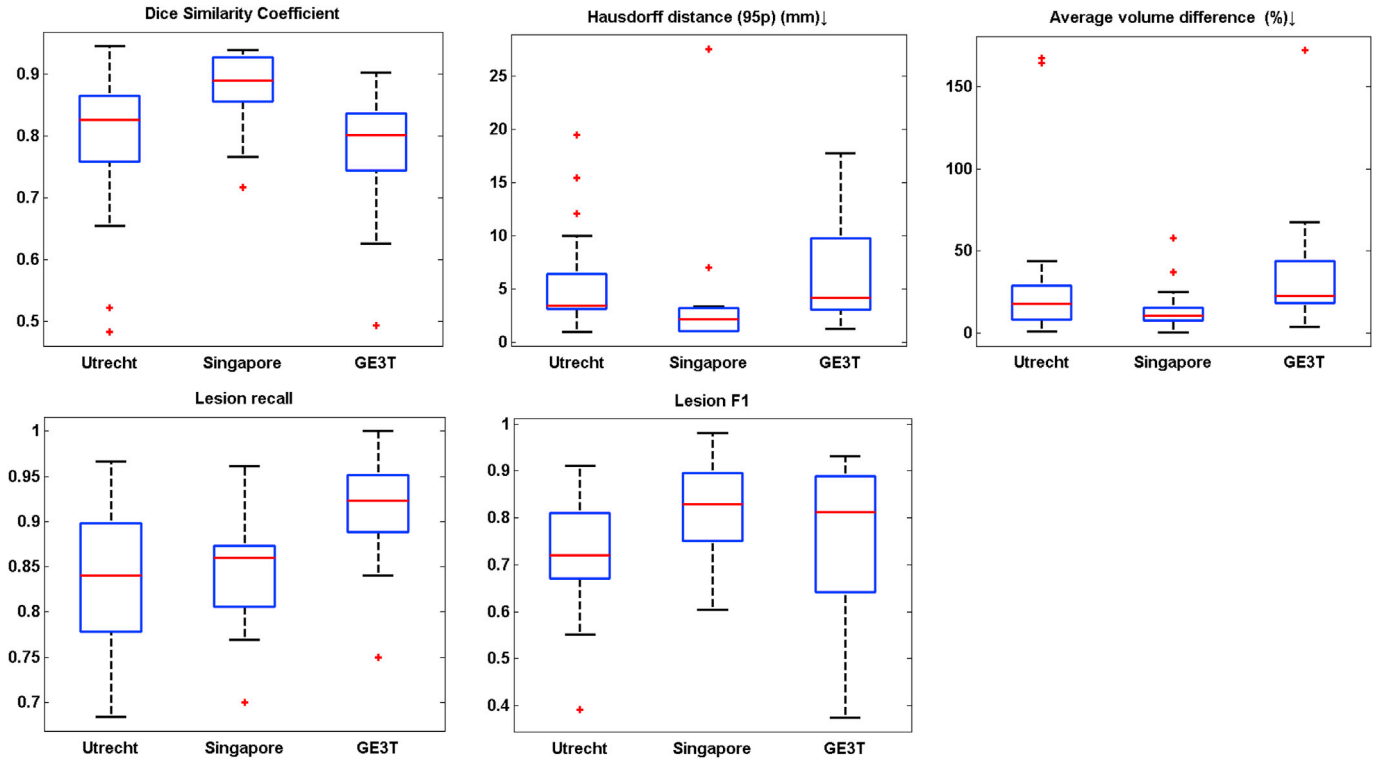


Fig. 7. Box plots of cross-scanner evaluation on the public training data. Each box plot summarizes the segmentation performance on subject from three testing scanners using one specific metric. For example, for box plot *Utrecht* in the upper left figure, it shows the distribution of segmentation results on *Utrecht* when training the model by using data from two other scanners - *Singapore* and *GE3T*.

5.1. Why choose 2D architecture

It is noted that there exist several 3D convolutional network architectures for brain tumor segmentation (Kamnitsas et al., 2017b; Havaei et al., 2017). The main motivation of employing 3D architectures is to extract rich spatial and contextual information from tumor/lesion tissue volume. However, in case of WMH segmentation, small lesions with high discontinuity and low contrast are commonly found, which contain poor spatial and contextual information. Furthermore, the imaging resolution along z-direction of the contest images is rather poor, and there exists large variation of spatial resolution as shown in Table 1, which further restricts the use of 3D deep learning models. Fig. 8 shows the case 11 in dataset *Utrecht*, in which small lesions with discontinuity characteristic are observed. Therefore a 2D architecture is chosen for this challenge to explore the texture information at slice level, and to drastically reduce the computational complexity. Data augmentation further equips the 2D model with desired invariance and robustness. It should be acknowledged that, when large clinical datasets are available in future, 3D architectures might help to improve the segmentation performance.

5.2. Analysis of U-Net hyper parameters

An appropriate parameter setting is crucial to successful training of deep fully convolutional networks. Here we mainly discuss some hyper parameters including the number of epochs, size of batch training and learning rate.

We selected the number of epochs for stopping training by contrasting training loss and validation loss over epochs. We split the public training dataset into a training set and a validation set by randomly picking 80% and the remaining 20% cases from each scanner respectively. Thus in total, the models were trained on 48 cases and validated on 12 cases. Fig. 9 shows the curves of training and validation loss over 100 epochs. It could be observed that the validation loss did not show a descending trend at around 50 epochs. The reason to choose 50 epochs rather than a higher one is 1) to avoid over fitting on the training data, and 2) keep low computational cost.

The size of batch and learning rate have a large influence on the stability of the training process. To our empirical observation, if the learning rate was set to values bigger than 10^{-3}, the training loss would be suddenly reaching to nearly 0 (i.e., the worst performance) at some beginning epoch and would remain not updating the training loss. Both of the batch size and learning rate directly influence the magnitude of the gradient and sometimes will lead to a gradient exposure issue. Therefore the batch size was set to 30 and learning rate was set to 0.0002 throughout all of the experiments.

5.3. Influence of imaging modalities

The T1 modality is known to provide a good contrast between the healthy tissues of the brain while FLAIR sequences are widely used to distinguish pathologies present in the white matter. Based on this, we assumed that these two modalities can provide complementary information for segmenting WMH. According to previous work (Dyrby et al., 2008), a combination of FLAIR and other modalities significantly

Table 6
Performance on two unseen scanners of top-5 methods among the 20 teams. The cells in gray shading indicate the best segmentation performance on each metric. The overall ranking is based on the average of the rank scores on each metric as shown in last row of Table 3. ↓ indicates that smaller value represents better performance.

Teams	Rank/score	DSC	H95↓	AVD↓	Recall	F1
Ours	**1/0.040**	0.745	11.04	26.2	0.87	0.725
nih_cidi_2	2/0.234	0.705	9.745	21.94	0.79	0.685
cian	3/0.264	0.745	14.10	28.425	0.82	0.665
nic – vicorob	4/0.374	0.715	13.53	56.31	0.815	0.62
nlp_logix	5/0.408	0.685	12.98	27.9	0.665	0.73

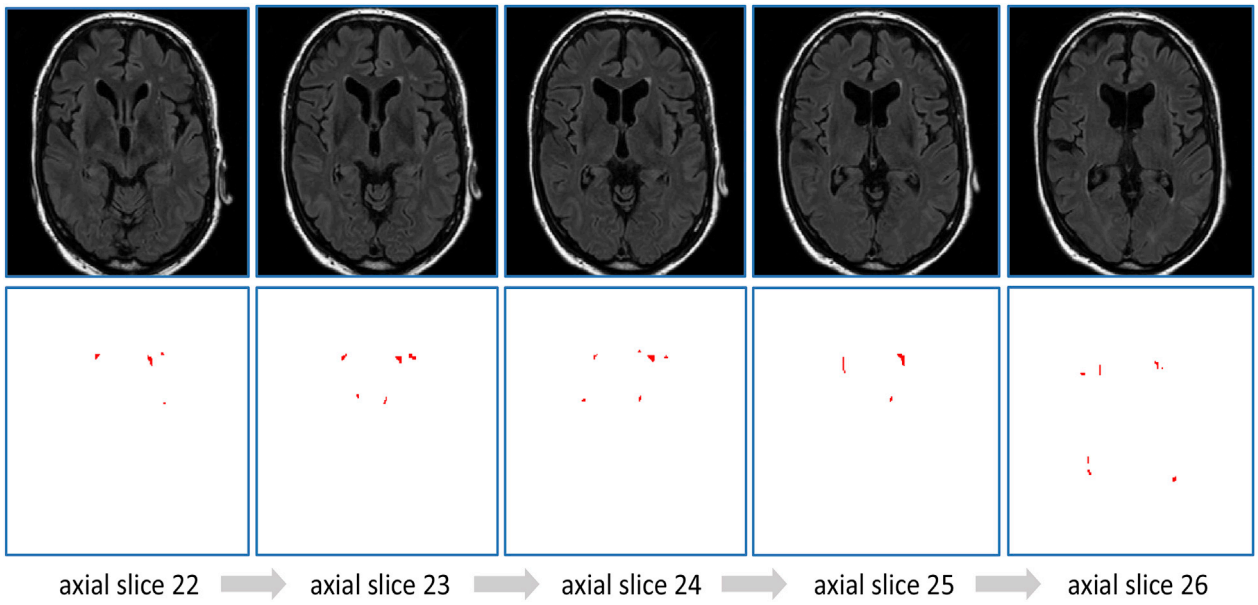


Fig. 8. Case 11 from the public training set shows the high discontinuity. From top to down, slices and corresponding ground-true segmentation maps. From left to right: axial slices from 22 to 26 and the corresponding ground truth.

improved the segmentation performance than using FLAIR alone. However, whether this combination improves the generalization capability to unseen scanner, has not been clearly investigated. We therefore analysed and presented a novel study for comparison in a cross-scanner-evaluation manner.

Table S2 to Table S4 in supplemental material report extensive results. They show that the combination of FLAIR and T1 slightly outperformed FLAIR alone on most of the metrics, suggesting T1 modality could provide useful information for detecting WMH. In Fig. 10 we showed the segmentation results of a case from *Singapore* tested by the model trained on *Utrecht* and *GE3T*. We observed that some false negatives were removed by using the combination of FLAIR

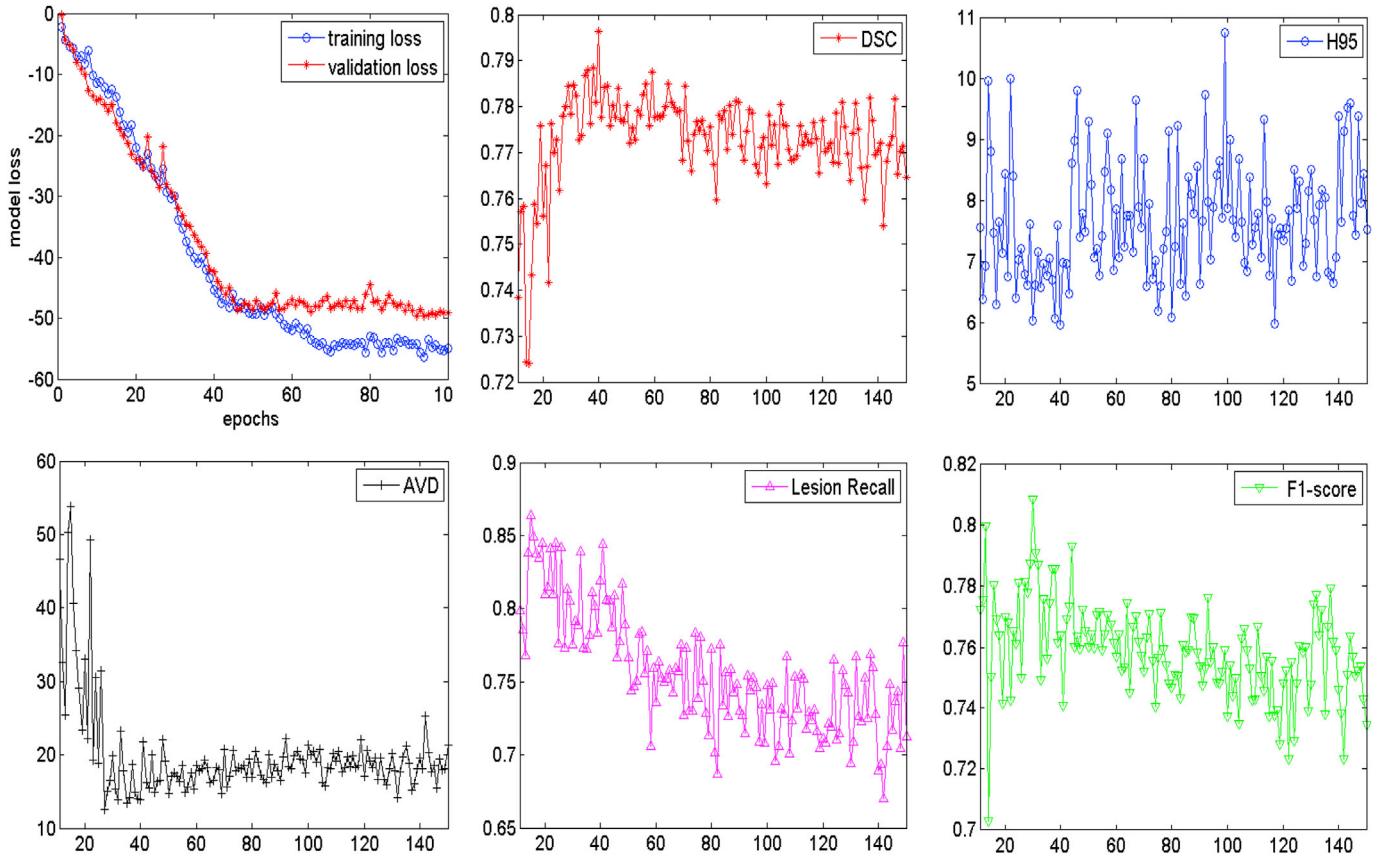


Fig. 9. Curves of training and validation loss and segmentation performance of each metric over epochs.

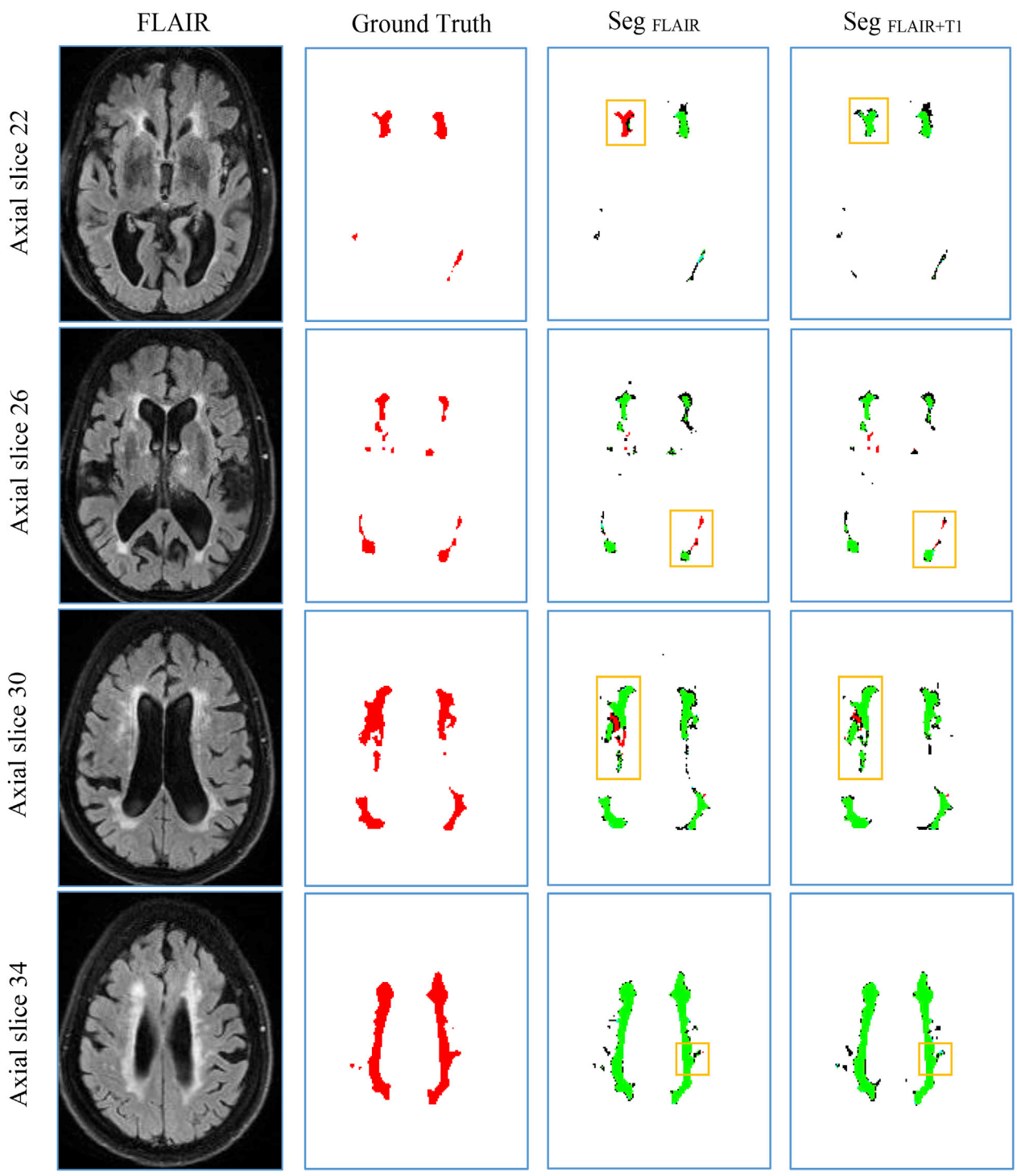


Fig. 10. Segmentation result on *Singapore 34*. From top to bottom: four axial slices of the same subject. From left to right: FLAIR MR images, the associated ground truth, segmentation result using FLAIR modality only and segmentation result using FLAIR and T1 modalities. In column $\text{Seg}_{\text{FLAIR}}$ and $\text{Seg}_{\text{FLAIR+T1}}$, the green area is the overlap between the segmentation maps and the ground-truth, the red pixels are the false negatives and the black ones are the false positives. (Best viewed in colour).

and T1 after comparing the column $\text{Seg}_{\text{FLAIR+T1}}$ and $\text{Seg}_{\text{FLAIR}}$, suggesting T1 provided complementary information on judging WMH. We further performed Wilcoxon signed rank test on the 60 cases. The improvements on H95 and F1-score were significant, giving p-values smaller than 1×10^{-4}.

5.4. Influence of data augmentation

The intention of data augmentation is generating training samples with different distributions to teach network learning desired invariance and robustness. We evaluated this technique using the cross-scanner evaluation as discussed in Section 5.3. The same experimental setting was used.

Table S5 to Table S7 in supplemental material report extensive results. They show that using data augmentation slightly improved segmentation results on most of the metrics. Fig. 11 shows the segmentation results of a case from *Utrecht* tested by the model trained on *Singapore* and *GE3T*. We observed that some false positives with small

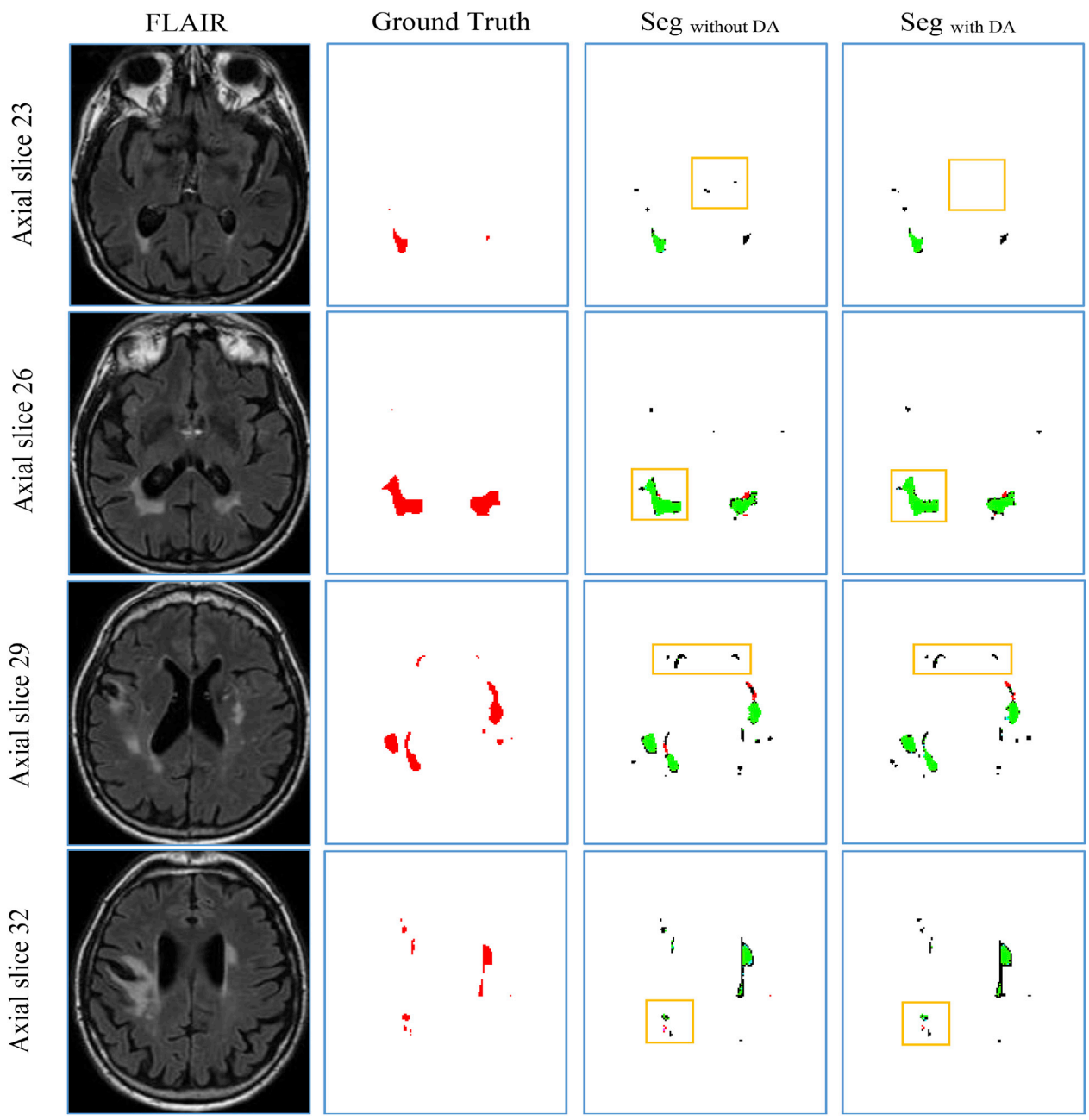


Fig. 11. Sample segmentation result on *Utrecht 04*. From top to bottom: four axial slices of the same subject. From left to right: FLAIR MR images, the associated ground truth, segmentation result without using data augmentation and segmentation result with data augmentation. In column $Seg_{withoutDA}$ and Seg_{withDA}, the green area is the overlap between the segmentation result and the ground truth, the red ones are the false negatives, and the black ones are the false positives. (Best viewed in colour).

volumes were removed by employing data augmentation after comparing the column $Seg_{withoutDA}$ to Seg_{withDA}, suggesting the model achieved robustness to small lesions. We further performed Wilcoxon signed rank test on the 60 cases. The improvements on H95, Recall and F1-score are statistically significant, giving p-values smaller than 1×10^{-4}.

5.5. Adaptability to different scanners

To ensure the usability of the proposed system in real world practice, which involves imaging data from various scanners and protocols, we evaluated its adaptability to imaging data across scanners. Extensive experiments were conducted by comparing the segmentation performances between models trained on either a single scanner or multiple ones.

Firstly, three sub-datasets from three scanners were evaluated independently. For example, 20 subjects from *Utrecht* were split into training set and test set, and each subject was evaluated using the leave-one-subject-out evaluation introduced in Section 4.2. Then the segmentation performance on each subject was compared to the one achieved by model trained on *additional* data from other two scanners. This comparison allows us to see the adaptability of the system.

Fig. 12 shows box plots of performances on each dataset. Interestingly, we observed that, on four metrics - *dice similarity coefficient, Hausdorff distance (95p), average volume difference and lesion F1-score*, the model trained on three scanners achieved significant improvement over the one trained on single scanner. However, on *lesion recall*, the model trained on single scanner gained slightly better segmentation

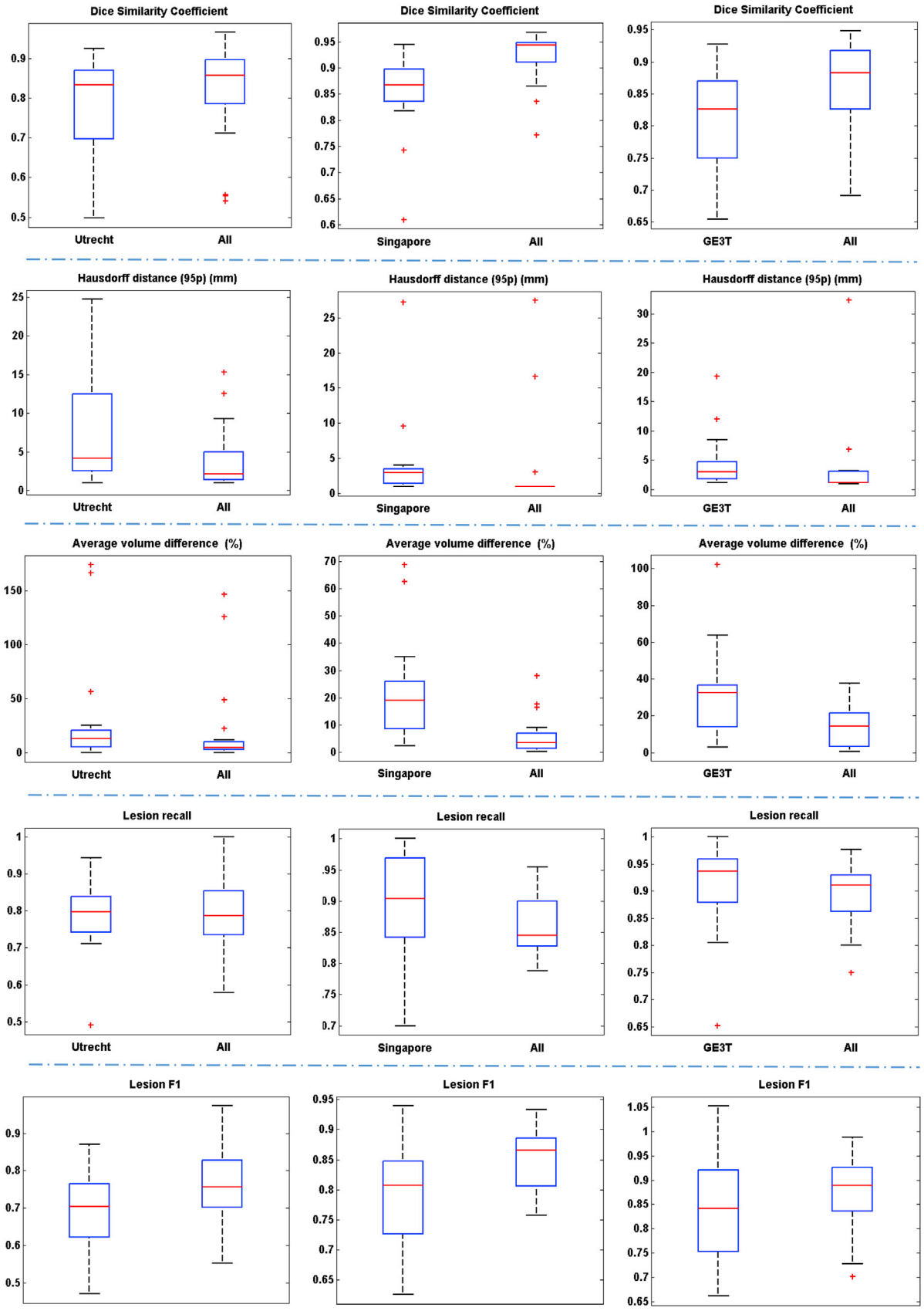


Fig. 12. Box plots of model adaptability evaluation. For example, the box plot in the left of first row shows two dice score distributions generated by two models trained on *Utrecht* only and *Utrecht* with additional data from other two scanners, respectively. From top to down: comparison of segmentation result on five metrics respectively. From left to right, comparison of segmentation result on *Utrecht*, *Singapore* and *GE3T* respectively.

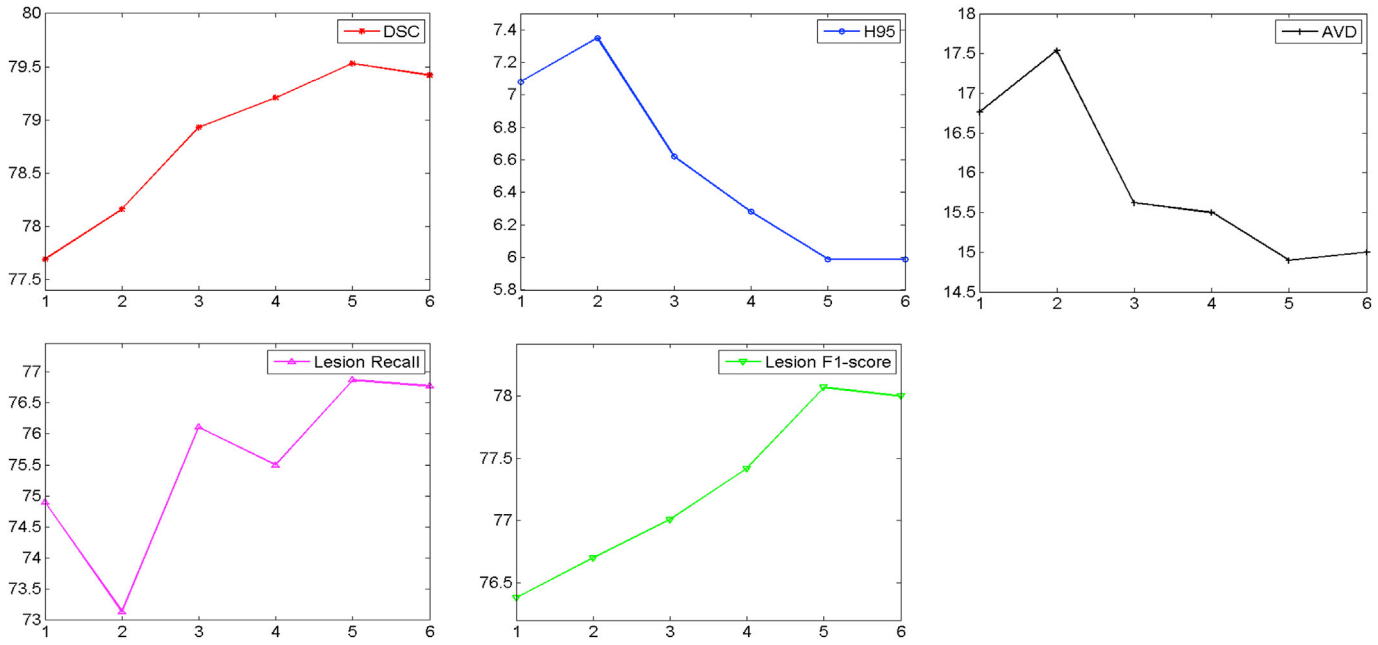


Fig. 13. Segmentation performance on validation set w. r.t ensemble size. The horizontal axis represents the number of models in the ensemble. We used an ensemble of three models in our final submission to the challenge.

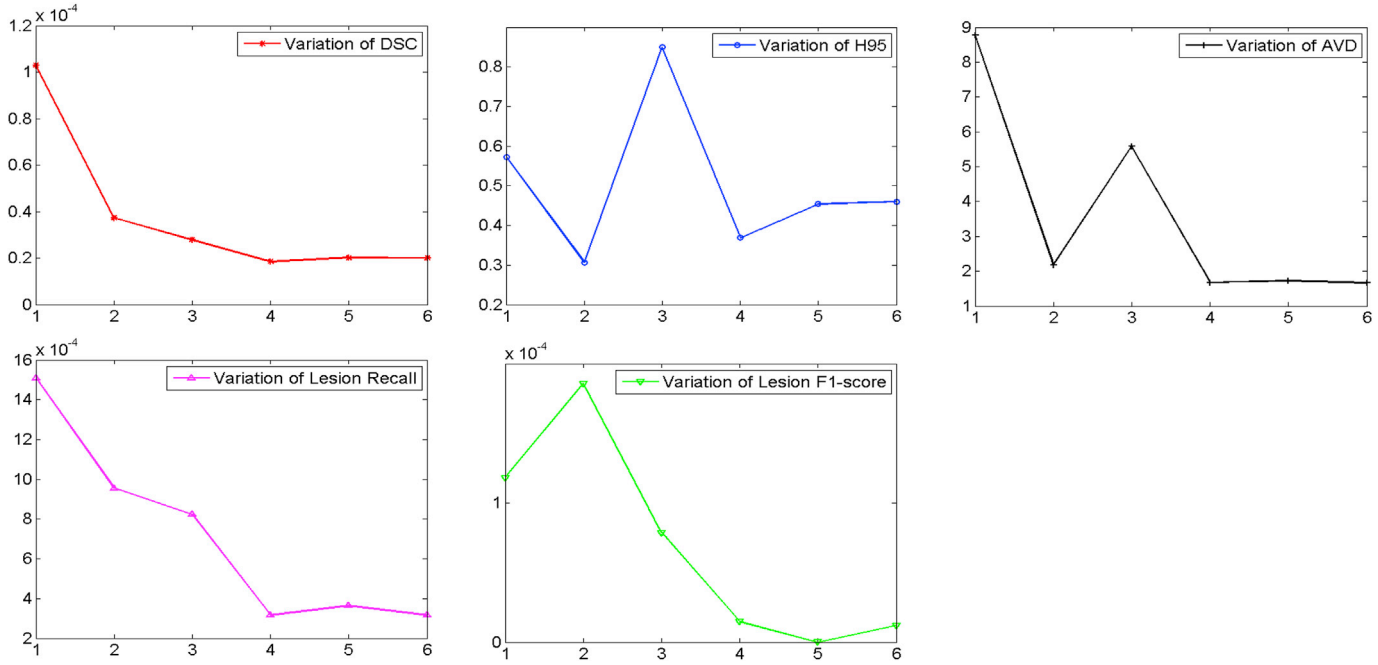


Fig. 14. The standard deviation of segmentation performance on validation set w. r.t ensemble size. We observed that the variation of segmentation performance was reduced when the size was increased.

performance. This was due to the decrease of the number of undetected small lesions. We concluded that the network trained on the larger data set that included cases obtained from different scanners shows better prediction performance, but at the cost of a sensitivity towards small lesions that were still detected best by networks trained on scanner- or sequence-specific data.

5.6. *Effect of the size of ensembles*

Ensemble learning aims at aggregating different models to boost the segmentation performance. The optimal size of an ensemble, i.e., how many models in the ensemble are needed, still remains an open issue and, as in many related ensemble learning task, a task specific parameter that needs to be optimized. To this end, we evaluated how the segmentation performance behaves over the number of ensemble models. We split the public dataset into training set and validation set by randomly picking 80% and 20% cases from each scanner respectively. The models were trained on 48 cases and validated on 12 cases. Then the segmentation performance on 12 cases were averaged on each evaluation metric. For each model with different size of ensembles, the training process was repeated five times and the segmentation results on the validation set were averaged.

Fig. 13 shows the curves of segmentation performance on five metrics w. r.t different ensemble size. It could be seen that (1) the ensemble with three or more models clearly outperformed the ensemble of only one model on all of the five metrics. The improvement of ensemble model with size 5 over one with size 3 is statistically significant on five metrics, all with small p-values; (2) when the size was further increased, performance tended to saturate and minor improvements in some of the measures came at the cost of small decreased in others. Fig. 14 shows standard deviation of segmentation performance between five repeated trained models w. r.t different ensemble size. It could be observed that the variation of segmentation performance was reduced on the main evaluation metrics when the size of ensemble was increased. It demonstrated that the ensemble model can not only boost the segmentation performance but also guarantee a robust segmentation result. Fig. 15 shows a case segmented by three individual models and their ensemble. We observed that three models trained with different weights initializations and shuffled data generated significantly different result on boundary and small lesions. And the model ensemble avoided the worst segmentation result.

5.7. Statistical analysis

5.7.1. Contribution of each component

We investigated in depth the contribution of each component using statistical analysis. Specifically, the performance of the proposed framework with and without a specific component was compared statistically as detailed below. For each of these comparisons, the public training dataset (from 60 patients) was first split into a training set and a validation set with a ratio of 4:1, resulting in a set of 48 training cases and a set of 12 validation cases. Then the proposed framework without a specific component was trained on the 48 training cases and evaluated on each of the 12 validation cases w. r.t each of the five organizer-provided evaluation metrics. The same protocol was also aplied to evaluate the complete proposed framework (i.e., without removing any component). Then for each metric, Wilcoxon signed rank test was adopted to test the statistical significance of the difference between the proposed framework *with* and *without* a specific component based on their validation performance. Since the comparisons were under a setting of multiple hypothesis testing, the p-values obtained for those five metrics were further adjusted by controlling the false discovery rate (PDR) for these hypothesis tests using the procedure proposed by Benjamini and Hochberg (1995). Table 7 summarizes the contributions of each component in the framework as well as PDR-adjusted p-values of the test. It could be observed that *preprocessing*, *data augmentation* and *ensemble model* have consistent improvements on all of the five metrics. In particular, all the improvements of using data augmentation show statistical significance with very small p-values. On two metrics (H95 and AVD), the improvements of preprocessing are statistically significant. Similarity, the use of ensemble improves the performances on all the five metrics, among which, three (DSC, H95, AVD) are statistically significant. The use of the two modalities improves the performances on four metrics although no improvement was observed on AVD metric.

Overall, the combination of these framework components helps build the state-of-the-art WMH segmentation system and differentiates our entry from other entries in the WMH segmentation competition.

5.7.2. Best-performing model vs ensemble model

In practise, compared to the use of the ensemble for testing, one

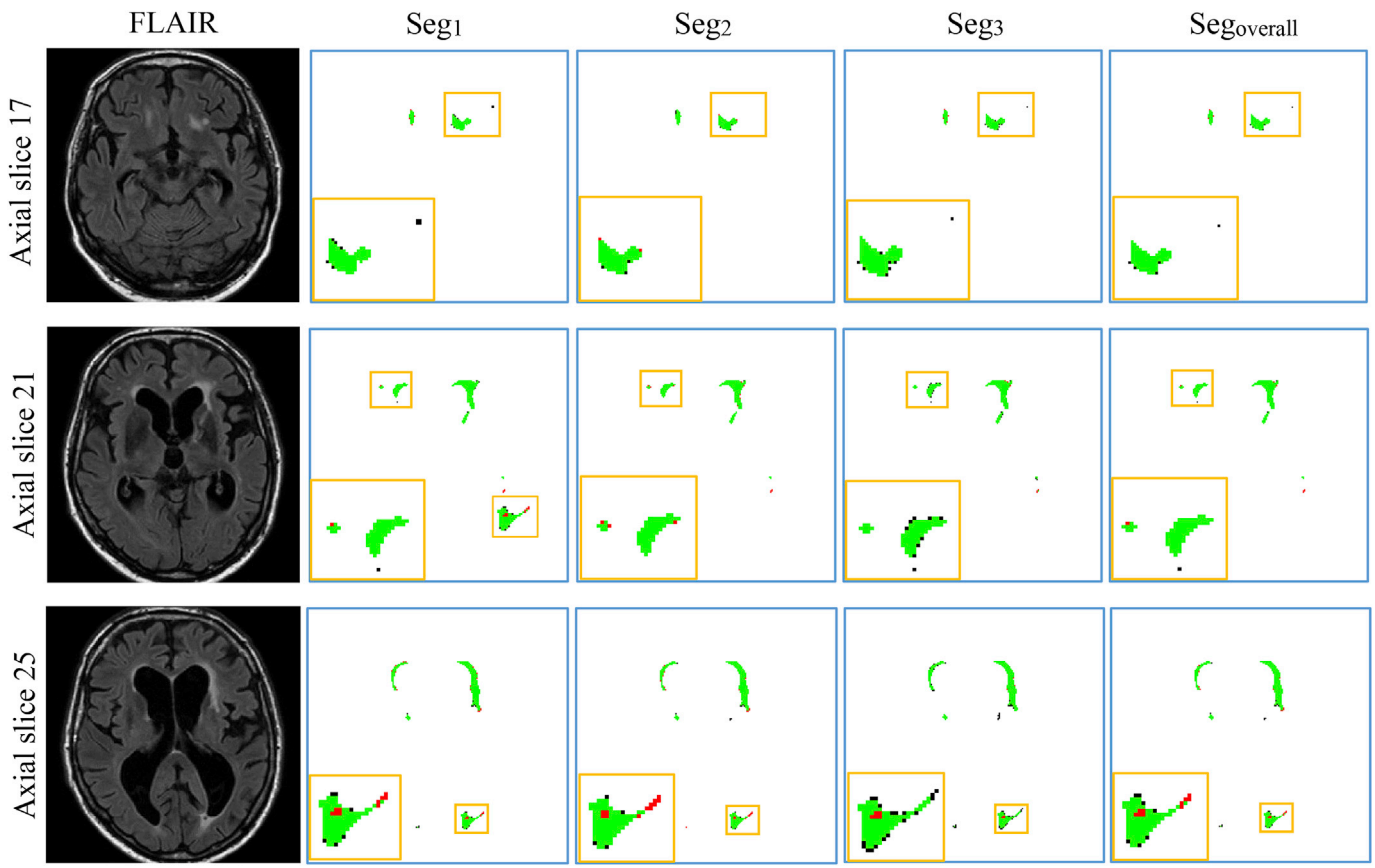


Fig. 15. Detailed segmentation results of three models and the ensemble. Columns Seg_1, Seg_2, Seg_3 and $Seg_{overall}$ represent the segmentation result generated by *model 1*, *model 2*, *model 3* and their *ensemble*. The green area in column Seg_1, Seg_2, Seg_3 and $Seg_{Overall}$ is the overlap between the segmentation result and ground truth. The red ones are the false negatives while the black ones are the false positives. For better visualization, the regions inside the smaller yellow bounding box are zoomed into the larger bounding box.

Table 7
The contribution of each component in the framework. *p-val* denotes the adjusted p-value after controlling false discovery rate, and its bold face indicates statistical significance. *IM* denotes the average improvement.

	Preprocess		Data Aug.		Modalities		Ensemble	
	IM	*p-val*	IM	*p-val*	IM	*p-val*	IM	*p-val*
DSC	1.04%	0.1067	1.38%	**0.0030**	0.62%	0.3393	1.98%	**0.0115**
H95 (mm)↓	0.2	**0.0013**	0.58	**0.0025**	0.57	**0.0013**	0.95	**0.0025**
AVD↓	2.15%	**0.0013**	3.02%	**0.0025**	−0.96%	0.0013	2.29%	**0.0025**
Recall	3.87%	0.1100	3.89%	**0.0425**	0.87%	0.4238	3.19%	0.9097
F1-score	4.11%	0.1100	5.72%	**0.0030**	1.70%	0.3766	1.70%	0.5871

alternative approach is to selected a model from the ensemble, which performs the best on the validation set as the candidate model for testing. We refer this model as a *best-performing* model. Here, we further compared the performances of best-performing model based on Dice loss and ensemble model. Specifically, the public training dataset (60 cases) was split into a training set, a validation set and a test set with a ratio of 3:1:1, resulting in 36 training cases, 12 validation cases and 12 test cases. We trained five models with different initializations, and selected the best-performing model based on the validation loss on the validation set. Then the performance of the best-performing model and the ensemble of the 5 models were compared on the *test* set. The averaged results on 12 test cases as well as the adjusted p-values of the Wilcoxon signed rank test after controlling the false discovery rate are shown in Table 8. It shows that ensemble model outperforms single best-performing model on four metrics (significantly on Dice score and lesion F1-score).

5.8. Computational complexity

All of the experiments were conducted on a GNU/Linux server running Ubuntu 16.04, with 32 GB RAM memory. The number of trainable parameters in the proposed model with two-channel inputs (FLAIR & T1) is 8,748,609. The algorithms were trained on a single NVIDIA Titan-Xp GPU with 12 GB RAM memory. It takes around 180 min to train a single model for 50 epochs on a training set containing 10,000 images of size 200 × 200 each. For testing, the segmentation of one scan with 48 slices by an ensemble of three models takes around 60 s using a Intel Xeon CPU (E3-1225v3) (without the use of GPU). In contrast, the segmentation per scan takes only 8 s when using a GPU.

6. Conclusions

In this paper we describe in detail our winning entry for MICCAI-2017 WMH Segmentation Challenge. To investigate the contribution of each component of our system, we empirically study the effects of imaging modalities and data augmentation as well as ensemble size used in the system training that all contributed to the performance of our segmentation model. We found that (1) FLAIR and T1 imaging modalities provide complementary information to judge WMH; (3) the proposed system shows good adaptability on various scanners and protocols; (4) ensemble model helps to reduce over-fitting and boost segmentation results. They are important factors to consider in building state-of-the-art WMH segmentation systems with good generalization capability. The methods employed by the top-5 teams in the challenge are all deep-learning models, suggesting deep-learning techniques especially convolutional networks show high efficacy in WMH segmentation. Although the segmentation results on 110 secret cases show its potential for real-world clinical use, the detection of small-volume WMH in MR images remains a challenging problem and is a future direction for the upcoming research in automated WMH segmentation. Some interesting architecture which learns context information between slices Chen et al. (2016) could be further investigated in future work. It will be interesting to discuss how segmentation difference between the algorithm and doctors will affect the clinical adoption, and how to address such a difference. This will need to test the algorithm in a clinical setting and get further feedback from radiologist and related therapist, which will be an interesting task in future work. Note that our brain intensities are normalized based on all of the voxels within the brain in order to calibrate intensities across scanners. Since patients have varying amount of (hyper-intense) diseases, which may bias the mean intensities used in the normalization. To alleviate this bias, robust measures can be used, such as robust mean or median absolute deviance. Alternatively, the lesion segmentation can be iterated and lesion areas identified in the first iteration are excluded in the normalization in the next iteration. We make our *Python* segmentation code available in *GitHub*.

Table 8
Comparison of the best-performing model and ensemble model. The adjusted p-values in bold indicate significant improvement achieved by ensemble model.

Models	DSC	H95 ↓	AVD ↓	Recall	F1
best-performing	77.06%	7.87 mm	16.78%	71.60%	72.99%
ensemble model	78.80%	7.18 mm	18.92%	72.66%	77.29%
improvement	1.74%	0.71 mm	−2.14%	0.84%	4.30%
p-value	**0.0015**	0.20	0.0772	0.1496	**0.0005**

Acknowledgment

We thank the MICCAI-2017 WMH Challenge organizer Dr. Hugo J. Kuijf and the joint effort of the UMC Utrecht, VU Amsterdam, and NUHS Singapore for making the datasets and test results available for this research. We thank Haocheng Shen for the discussion during the challenge and presentation in the MICCAI BrainLes Workshop. The work was initialised when Hongwei Li was a visiting student in University of Dundee. The WMH contest submission was completed by a joint team from CVIP, Computing in University of Dundee and Sun Yat-Sen University. This work was supported in part by NSFC grant (No. 61628212), China; Royal Society International Exchanges grant (No. 170168), UK; the Macao Science and Technology Development Fund under 112/2014/A3, China and Guangdong Program (No. 2014B010118003) China.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.neuroimage.2018.07.005.

References

Anbeek, P., Vincken, K.L., Van Osch, M.J., Bisschops, R.H., Van Der Grond, J., 2004. Probabilistic segmentation of white matter lesions in MR imaging. Neuroimage 21 (3), 1037–1044.

Beare, R., Srikanth, V., Chen, J., Phan, T.G., Stapleton, J., Lipshut, R., Reutens, D., 2009. Development and validation of morphological segmentation of age-related cerebral white matter hyperintensities. Neuroimage 47 (1), 199–203.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Stat. Soc. B 57, 289–300.

Chen, J., Yang, L., Zhang, Y., Alber, M., Chen, D.Z., 2016. Combining fully convolutional and recurrent neural networks for 3d biomedical image segmentation. In: Advances in Neural Information Processing Systems, pp. 3036–3044.
Debette, S., Markus, H., 2010. The clinical importance of white matter hyperintensities on brain magnetic resonance imaging: systematic review and meta-analysis. BMJ 341, c3666.
Dyrby, T.B., Rostrup, E., Baaré, W.F., van Straaten, E.C., Barkhof, F., Vrenken, H., Ropele, S., Schmidt, R., Erkinjuntti, T., Wahlund, L.-O., et al., 2008. Segmentation of age-related white matter changes in a clinical multi-center study. Neuroimage 41 (2), 335–345.
Geremia, E., Menze, B.H., Clatz, O., Konukoglu, E., Criminisi, A., Ayache, N., 2010. Spatial decision forests for MS lesion segmentation in multi-channel MR images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 111–118.
Geremia, E., Clatz, O., Menze, B.H., Konukoglu, E., Criminisi, A., Ayache, N., 2011. Spatial decision forests for MS lesion segmentation in multi-channel magnetic resonance images. Neuroimage 57 (2), 378–390.
Ghafoorian, M., Karssemeijer, N., Heskes, T., Uden, I.W., Sanchez, C.I., Litjens, G., Leeuw, F.-E., Ginneken, B., Marchiori, E., Platel, B., 2017. Location sensitive deep convolutional neural networks for segmentation of white matter hyperintensities. Sci. Rep. 7 (1), 5110.
Gouw, A.A., Seewann, A., Van Der Flier, W.M., Barkhof, F., Rozemuller, A.M., Scheltens, P., Geurts, J.J., 2010. Heterogeneity of small vessel disease: a systematic review of MRI and histopathology correlations. J. Neurol. Neurosurg. Psychiatr. jnnp–2009.
Grimaud, J., Lai, M., Thorpe, J., Adeleine, P., Wang, L., Barker, G., Plummer, D., Tofts, P., McDonald, W., Miller, D., 1996. Quantification of MRI lesion load in multiple sclerosis: a comparison of three computer-assisted techniques. Magn. Reson. Imag. 14 (5), 495–505.
Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.-M., Larochelle, H., 2017. Brain tumor segmentation with deep neural networks. Med. Image Anal. 35, 18–31.
He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.
Russakovsky, Olga, Deng, Jia, Su, Hao, Krause, Jonathan, Satheesh, Sanjeev, Ma, Sean, Huang, Zhiheng, Karpathy, Andrej, Khosla, Aditya, Bernstein, Michael, Berg, Alexander C., Fei-Fei, Li, 2015. Imagenet large scale visual recognition challenge. I. J. Com. 211–252. Vision 115.3.
Kamnitsas, K., Bai, W., Ferrante, E., McDonagh, S., Sinclair, M., Pawlowski, N., Rajchl, M., Lee, M., Kainz, B., Rueckert, D., et al., 2017a. In: Ensembles of Multiple Models and Architectures for Robust Brain Tumour Segmentation arXiv preprint arXiv: 1711.01468.
Kamnitsas, K., Ledig, C., Newcombe, V.F., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B., 2017b. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Med. Image Anal. 36, 61–78.
Kim, K.W., MacFall, J.R., Payne, M.E., 2008. Classification of white matter lesions on magnetic resonance imaging in elderly persons. Biol. Psychiatr. 64 (4), 273–280.
Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105.
Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440.
Maier, O., Menze, B.H., von der Gablentz, J., Häni, L., Heinrich, M.P., Liebrand, M., Winzeck, S., Basit, A., Bentley, P., Chen, L., et al., 2017. ISLES 2015-A public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI. Med. Image Anal. 35, 250–269.
Medical Image Computing and Computer-Assisted Intervention−MICCAI, 2017. In: Descoteaux, Maxime, Maier-Hein, Lena, Franz, Alfred, Jannin, Pierre, Collins, D. Louis, Duchesne, Simon (Eds.), 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Vol. 10435. Springer, 2017. Lecture Notes in Computer Science.
Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al., 2015. The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans. Med. Imag. 34 (10), 1993–2024.
Merkel, D., 2014. Docker: lightweight linux containers for consistent development and deployment. Linux J. 239, 2 (2014).
Milletari, F., Navab, N., Ahmadi, S.-A., 2016. V-net: fully convolutional neural networks for volumetric medical image segmentation. In: 3D Vision (3DV), 2016 Fourth International Conference on, IEEE, pp. 565–571.
Moeskops, P., de Bresser, J., Kuijf, H.J., Mendrik, A.M., Biessels, G.J., Pluim, J.P., Išgum, I., 2017. Evaluation of a Deep Learning Approach for the Segmentation of Brain Tissues and white Matter Hyperintensities of Presumed Vascular Origin in MRI. Clinical, NeuroImage.
Opitz, D.W., Maclin, R., 1999. Popular ensemble methods: an empirical study. J. Artif. Intell. Res. 11, 169–198.
Pantoni, L., 2010. Cerebral small vessel disease: from pathogenesis and clinical characteristics to therapeutic challenges. Lancet Neurol. 9 (7), 689–701.
Peng, C., Zhang, X., Yu, G., Luo, G., Sun, J., 2017. Large Kernel Matters–improve Semantic Segmentation by Global Convolutional Network arXiv preprint arXiv:1703.02719.
Ronneberger, O., Fischer, P., Brox, T., U-net, 2015. Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.
Schmidt, P., Mühlau, M., Gaser, C., Wink, L., 2013. LST: a Lesion Segmentation Tool for SPM.
Simões, R., Mönninghoff, C., Dlugaj, M., Weimar, C., Wanke, I., van Walsum, A.-M. v. C., Slump, C., 2013. Automatic segmentation of cerebral white matter hyperintensities using only 3D FLAIR images. Magn. Reson. Imag. 31 (7), 1182–1189.
Simonyan, K., Zisserman, A., 2014. Very Deep Convolutional Networks for Large-scale Image Recognition arXiv preprint arXiv:1409.1556.
Smith, S.M., 2002. Fast robust automated brain extraction. Hum. Brain Mapp. 17 (3), 143–155.
Styner, M., Lee, J., Chin, B., Chin, M., Commowick, O., Tran, H., Markovic-Plese, S., Jewells, V., Warfield, S., 2008. 3D segmentation in the clinic: a grand challenge II: MS lesion segmentation. Midas Journal 1–6, 2008.
Van Leemput, K., Maes, F., Vandermeulen, D., Colchester, A., Suetens, P., 2001. Automated segmentation of multiple sclerosis lesions by model outlier detection. IEEE Trans. Med. Imag. 20 (8), 677–688.
Yoo, B.I., Lee, J.J., Han, J.W., Lee, E.Y., MacFall, J.R., Payne, M.E., Kim, T.H., Kim, J.H., Kim, K.W., et al., 2014. Application of variable threshold intensity to segmentation for white matter hyperintensities in fluid attenuated inversion recovery magnetic resonance images. Neuroradiology 56 (4), 265–281.

Received: 25 November 2020 | Revised: 23 August 2021 | Accepted: 25 August 2021

DOI: 10.1002/hbm.25655

RESEARCH ARTICLE

WILEY

Automated claustrum segmentation in human brain MRI using deep learning

Hongwei Li[1,2] | Aurore Menegaux[3,4] | Benita Schmitz-Koep[3,4] | Antonia Neubauer[3,4] | Felix J. B. Bäuerlein[5] | Suprosanna Shit[1] | Christian Sorg[3,4,6] | Bjoern Menze[1,2] | Dennis Hedderich[3,4]

[1]Department of Informatics, Technical University of Munich, Munich, Germany

[2]Department of Quantitative Biomedicine, University of Zurich, Zurich, Switzerland

[3]TUM-NIC Neuroimaging Center, Munich, Germany

[4]Department of Neuroradiology, Klinikum rechts der Isar, School of Medicine, Technical University of Munich, Munich, Germany

[5]Department of Molecular Structural Biology, Max Planck Institute of Biochemistry, Munich, Germany

[6]Department of Psychiatry, Klinikum rechts der Isar, School of Medicine, Technical University of Munich, Munich, Germany

Correspondence
Hongwei Li, IMETUM, Boltzmannstrasse 11, Garching, 85748, Bavaria, Germany.
Email: hongwei.li@tum.de

Funding information
German Federal Ministry of Education and Science, Grant/Award Numbers: BMBF 01ER0801, BMBF 01ER0803; Technische Universität München, Grant/Award Number: KKF 8765162; Federal Ministry of Education and Science, Grant/Award Numbers: BMBF 01ER0803, BMBF 01ER0801; Deutsche Forschungsgemeinschaft, Grant/Award Number: SO 1336/1-1

Abstract

In the last two decades, neuroscience has produced intriguing evidence for a central role of the claustrum in mammalian forebrain structure and function. However, relatively few in vivo studies of the claustrum exist in humans. A reason for this may be the delicate and sheet-like structure of the claustrum lying between the insular cortex and the putamen, which makes it not amenable to conventional segmentation methods. Recently, Deep Learning (DL) based approaches have been successfully introduced for automated segmentation of complex, subcortical brain structures. In the following, we present a multi-view DL-based approach to segment the claustrum in T1-weighted MRI scans. We trained and evaluated the proposed method in 181 individuals, using bilateral manual claustrum annotations by an expert neuroradiologist as reference standard. Cross-validation experiments yielded median volumetric similarity, robust Hausdorff distance, and Dice score of 93.3%, 1.41 mm, and 71.8%, respectively, representing equal or superior segmentation performance compared to human intra-rater reliability. The leave-one-scanner-out evaluation showed good transferability of the algorithm to images from unseen scanners at slightly inferior performance. Furthermore, we found that DL-based claustrum segmentation benefits from multi-view information and requires a sample size of around 75 MRI scans in the training set. We conclude that the developed algorithm allows for robust automated claustrum segmentation and thus yields considerable potential for facilitating MRI-based research of the human claustrum. The software and models of our method are made publicly available.

KEYWORDS
claustrum, deep learning, image segmentation, MRI, multi-view

1 | INTRODUCTION

The claustrum is a highly conserved gray matter structure of the mammalian forebrain, situated in the white matter between the putamen and the insular cortex, more precisely between the external and the extreme capsule (Kowiański, Dziewiątkowski, Kowiańska, & Moryś, 1999; Puelles, 2014). Although first described by Félix Vicq d'Azyr in the late 18th century, it has remained one of the most enigmatic structures of the brain (Johnson & Fenske, 2014). In a seminal article by Sir Francis Crick and Christof Koch, they proposed a role of the claustrum for processes that give rise to integrated conscious percepts (Crick & Koch, 2005), which has spurred new interest in the claustrum and its putative function. From animal and human studies, we know that the claustrum is the most widely connected gray matter structure in the brain in relation to its size, being connected to both the ipsilateral and the contralateral hemisphere (Mathur, 2014; Pearson, Brodal, Gatter, & Powell, 1982; Reser et al., 2017; Torgerson, Irimia, Goh, & Van Horn, 2015; Zingg et al., 2014; Zingg, Dong, Tao, & Zhang, 2018). It is reciprocally connected to almost all cortical regions including motor and somatosensory as well as visual, limbic, auditory, associative, and prefrontal cortices, and receives neuromodulatory input from subcortical structures (Goll, Atlan, & Citri, 2015; Torgerson et al., 2015). While the claustrum's exact function remains elusive, recent evidence suggests a role in basic cognitive processes such as selective attention or task switching (Brown et al., 2017; Mathur, 2014; Remedios, Logothetis, & Kayser, 2010, 2014). A rather new but equally interesting perspective on the claustrum is its unique ontogeny and a link to so-called subplate neurons, which have been proposed to play a role in neurodevelopmental disorders such as schizophrenia, autism, and preterm birth (Bruguier et al., 2020; Hoerder-Suabedissen & Molnár, 2015; Watson & Puelles, 2017).

Most human in vivo studies using MRI to investigate the claustrum suffer from small sample sizes (Arrigo et al., 2017; Krimmel et al., 2019; Milardi et al., 2015) since the sheet-like and delicate anatomy of the claustrum precludes classic atlas-based segmentation methods and is challenging for statistical shape models and traditional machine learning methods (Aljabar, Wolz, & Rueckert, 2012; Heimann & Meinzer, 2009). Consequently, manual annotation has typically been necessary, which is notoriously time-consuming, requires expert knowledge, and is not feasible to be applied in large-scale studies of the human brain (Arrigo et al., 2017; Milardi et al., 2015; Torgerson & Van Horn, 2014).

Thus, to promote our understanding of the claustrum in humans, an objective and accurate, automated, MRI-based segmentation method is needed. As mentioned before, the claustrum is not included as a region of interest (ROI) in most MR-based anatomic atlases of the brain. In fact, only BrainSuite, which is a tool for automated cortical parcellation and subcortical segmentation based on surface-constrained volumetric registration of individual MR images of the brain to a manually labeled atlas, contains the claustrum as a ROI (Joshi, Shattuck, Thompson, & Leahy, 2007). However, this method has been shown to be rather unreliable, most likely due to the challenging anatomy of the claustrum (Berman, Schurr, Atlan, Citri, & Mezer, 2020). Very recently, an automated, rule-based method using anatomical landmarks for claustrum segmentation has been published and showed improved segmentation accuracy compared to BrainSuite but still only less accuracy in comparison with manual annotations (Berman et al., 2020). In conclusion, there is still the need for improved fast and reproducible, automated segmentation of the claustrum in order to enable its exploration in large MRI studies.

In recent years, computer vision and machine learning techniques have been increasingly used in the medical field, pushing the limits of segmentation methods relying on atlases, statistical shape models and traditional machine learning approaches (Aljabar et al., 2012; Aljabar, Heckemann, Hammers, Hajnal, & Rueckert, 2009; Heimann & Meinzer, 2009). Particularly, deep learning (DL) (LeCun, Bengio, & Hinton, 2015) based approaches have shown promising results on various medical image segmentation tasks, for example, brain structure and tumor segmentation in MR images (Chen, Dou, Yu, Qin, & Heng, 2018; Kamnitsas et al., 2017; Prados et al., 2017; Wachinger, Reuter, & Klein, 2018). Recent segmentation methods commonly rely on so-called convolutional neural networks (CNNs). Applied to segmentation tasks, these networks "learn" proper structural information from a set of manually labeled data serving as ground truth for training. In the testing stage, CNNs perform automated segmentation on unseen images yielding rather high accuracies even for tiny structures such as white-matter lesions (Li et al., 2018). Recently, a clustering-based approach was proposed to segment the dorsal claustrum (Berman et al., 2020) and achieved <60% Dice coefficient when compared with manual segmentations. Yet DL-based approaches, which leverage large-scale datasets, have not been explored and can potentially improve segmentation accuracy.

Thus, we hypothesized that DL-based techniques used to segment the claustrum can fill the existing gap. Based on a large number of manually annotated, T1-weighted brain MRI scans, we propose a 2D multi-view framework for fully automated claustrum segmentation. In order to assess our central hypothesis, we will evaluate the segmentation accuracy of our algorithm on an annotated dataset using three canonical evaluation metrics and compare it to intrarater variability. Further, we will investigate whether multi-view information significantly improves the segmentation performance. In addition, we will address the questions of robustness against scanner type and how increasing the training set affects segmentation accuracy. To foreshadow results, we found robust, reliable, and stable claustrum segmentation based on our DL algorithm, which we make publicly available using an open-source repository: https://github.com/hongweilibran/claustrum_multi_view.

2 | MATERIALS AND METHODS

2.1 | Datasets

In the following two sections, we describe the datasets and evaluation metrics used in this study. T1-weighted three-dimensional scans of

181 individuals without known brain injury were included from the Bavarian Longitudinal Study (Hedderich et al., 2019). The study was carried out following the *Declaration of Helsinki* and was approved by the local institutional review boards. Written consent was obtained from all participants. The MRI acquisition took place at two sites: The Department of Neuroradiology, Klinikum rechts der Isar, Technische Universität München ($n = 120$) and the Department of Radiology, University Hospital of Bonn ($n = 61$). MRI examinations were performed at both sites on either a *Philips Achieva 3 T* or a *Philips Ingenia 3 T* system using 8-channel SENSE head-coils.

The imaging protocol includes a high-resolution T1-weighted, 3D-MPRAGE sequence (TI = 1300 ms, TR = 7.7 ms, TE = 3.9 ms, flip angle 15°; field of view: 256 mm × 256 mm) with a reconstructed isotropic voxel size of 1 mm^3. All images were visually inspected for artifacts and gross brain lesions that could potentially impair manual claustrum segmentation (Table 1).

2.2 | Preprocessing

Before manual segmentation, the images are skull-stripped using *ROBEX* (Iglesias, Liu, Thompson, & Tu, 2011) and denoised with spatially adaptive nonlocal means (Manjón, Coupé, Martí-Bonmatí, Collins, & Robles, 2010) to enhance the visibility of the claustrum. Manual annotations were performed by a neuroradiologist (D.M.H.) with 7 years of experience using a modified segmentation protocol (Davis, 2008) in ITK-SNAP 3.6.0 (Yushkevich et al., 2006). In brief, the claustrum was segmented in axial and coronal orientations, including its dorsal and ventral division at individually defined optimal image contrast for differentiation of gray and white matter. First, the claustrum was delineated on axial slices at the basal ganglia level, where it is visible continuously. Second, the claustrum was traced inferiorly until it was no longer visible. Consecutively, the claustrum was traced superiorly until its superior border. Notably, the superior parts of the claustrum are usually discontinuous below the insular cortex. Then, the axial annotations were checked and corrected (if necessary) using coronal views. This process is essential for the claustrum parts extending below the putamen and the ventral claustrum extending to the stem of the temporal lobe.

An additional preprocessing step is performed on top of the basic preprocessing steps carried out by the rater. We aim to normalize the voxel intensities to reduce the variations across subjects and scanners. Thus, a simple yet effective preprocessing step is used in both training and testing stages. It includes two steps: (1) cropping or padding each slice to a uniform size and (2) *z-score* normalization of the brain voxel intensities. First, all the axial and coronal slices are automatically cropped or padded to 180 × 180 to guarantee a uniform input size for the deep-learning model. Next, z-score normalization is performed for individual 3D scans. The mean and standard deviation are calculated based on the intensities within each individual's brain mask. Finally, the voxel intensities are rescaled to a mean of zero and unit standard deviation (Figure 1).

2.3 | Multi-view fully convolutional neural networks

2.3.1 | Multi-view learning

When performing manual annotations, neuroradiologists rely on axial and coronal views to identify the structure. Thus, we hypothesized that the image features from the two geometric views would be complementary to locate the claustrum and would be beneficial for reducing false positives on individual views. We train two deep CNN models on 2D single-view slices after parsing a 3D MRI volume into axial and coronal views. The sagittal view is excluded because we find it does not improve segmentation results. Further discussion is provided in Section 3.2. We propose a practical and straightforward approach to aggregate the multi-view information in probability space at a voxel-wise level during the inference stage (see Figure 2a). We train two single-view models on the 2D image slices from axial and coronal views, respectively. During the testing stage, we predict the single-view segmentation mask and fuse the multi-view information by averaging the voxel-wise probabilities.

2.3.2 | Single-view 2D convolutional network architecture

We built a 2D architecture based on a recent U-shape network (Li et al., 2018; Ronneberger, Fischer, & Brox, 2015) and tailored it for the claustrum segmentation task. The network architecture is delineated in Figure 2. It consists of a down-convolutional part that shrinks the spatial dimensions (left side), and an up-convolutional part that expands the score maps (right side). Skip connections between down-convolutional and up-convolutional are used. In this model, two convolutional layers are repeatedly employed, followed by a rectified

TABLE 1 Characteristics of the dataset in this study

Datasets	Scanner name	Voxel size (mm^3)	Number of subjects
Bonn-1	Philips Achieva 3 T	1.00 × 1.00 × 1.00	15
Bonn-2	Philips Ingenia 3 T	1.00 × 1.00 × 1.00	46
Munich-1	Philips Achieva 3 T	1.00 × 1.00 × 1.00	103
Munich-2	Philips Ingenia 3 T	1.00 × 1.00 × 1.00	17

Note: The dataset consists of 181 subjects from four scanners and two centers.

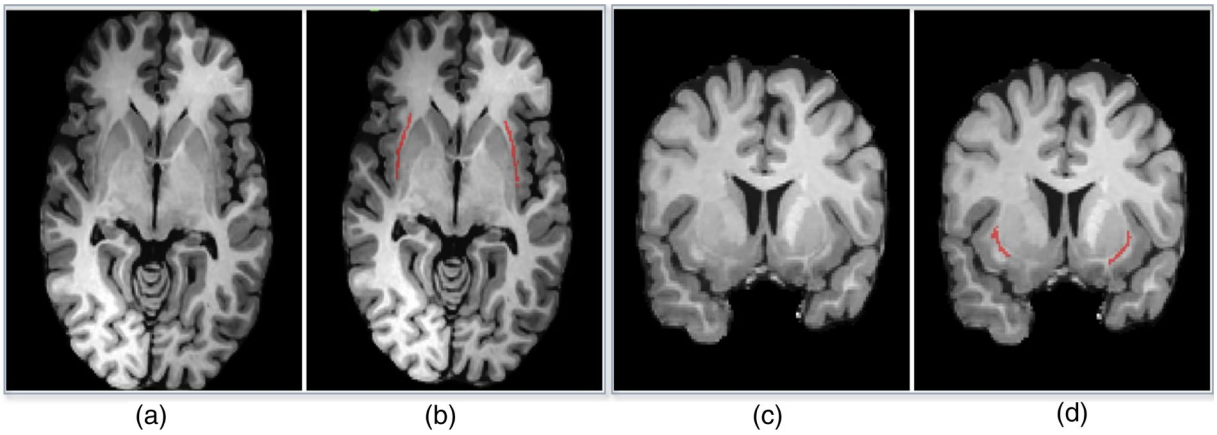


FIGURE 1 Examples of axial (a, b) and coronal (c, d) MR slices with corresponding manual annotation of the claustrum structure (in b and d) by a neuroradiologist

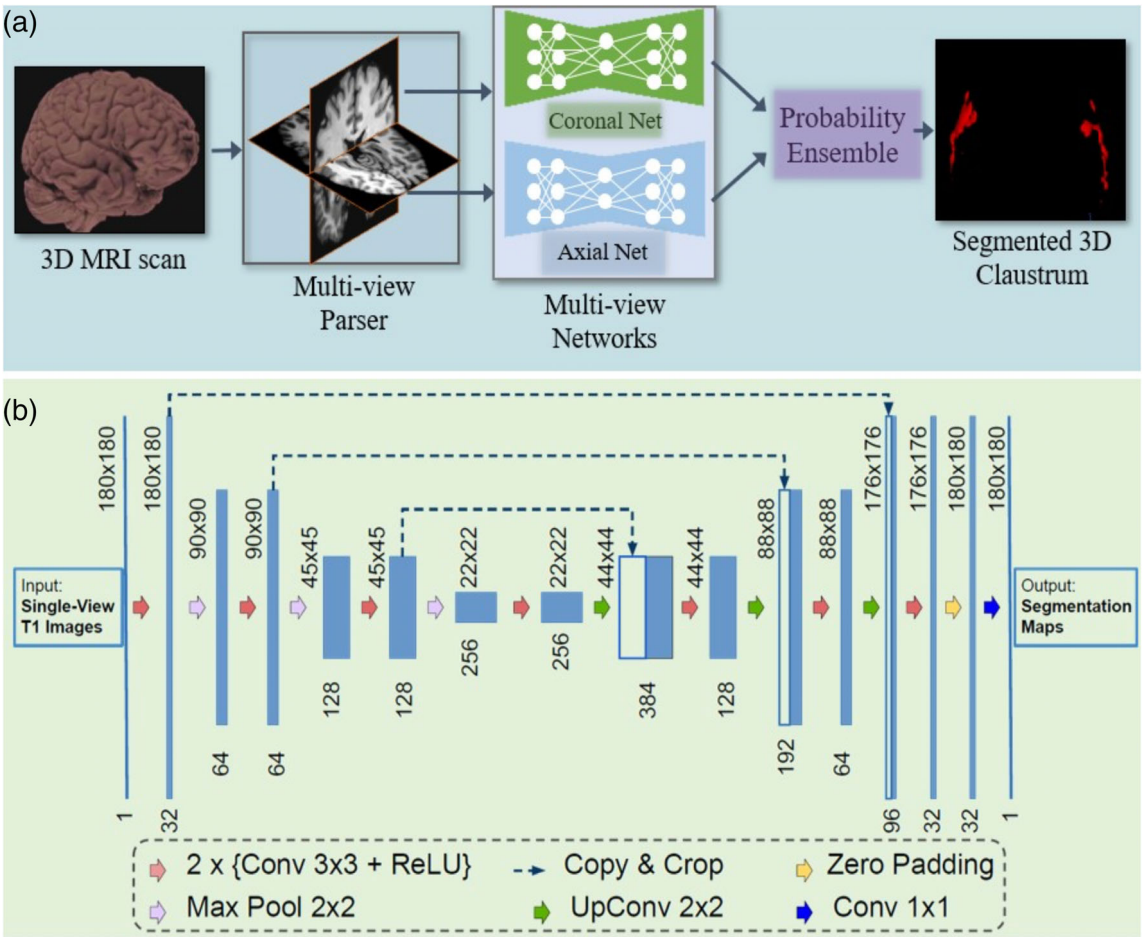


FIGURE 2 (a) A schematic view of the proposed segmentation method using multi-view fully convolutional networks to segment the 3D claustrum jointly; (b) 2D Convolutional network architecture for each view (i.e., axial and coronal). It takes the raw images as input and predicts its segmentation maps. The network consists of several nonlinear computational layers in a shrinking part (left side) and an expansive part (right side) to extract semantic features of the claustrum structure

linear unit (*ReLU*) and a 2 × 2 max pooling operation with stride 2 for down-sampling. At the final layer, a 1 × 1 convolution is used to map each 64-component feature vector to two classes. In total, the network contains 16 convolutional layers. The network takes the single-view slices of T1 modality scans as the input during training and testing (see Figure 2b).

2.4 | Loss function

With respect to the claustrum segmentation task, the numbers of positives (claustrum) and negatives (nonclaustrum) are highly unbalanced. One promising solution to tackle this issue is to use Dice loss (Milletari, Navab, & Ahmadi, 2016) as the loss function for training the model. The formulation is as follows.

Let $G = \{g_1,...,g_N\}$ be the ground-truth segmentation maps over N slices, and $P = \{p_1,...,p_N\}$ be the predicted probabilistic maps over N slices. The Dice loss function can be expressed as:

$$\text{Dice Loss} = -\frac{2\sum_{n=1}^{N}|p_n \circ g_n| + s}{\sum_{n=1}^{N}(|p_n| + |g_n|) + s}$$

where $\circ$ represents the entrywise product of two matrices, and $|\cdot|$ represents the sum of the matrix entries. The s term is used here to ensure the loss function stability by avoiding the division by 0, that is, in a case where the entries of G and P are all zeros. s is set to 1 in our experiments.

2.5 | Anatomically consistent postprocessing

The postprocessing for the 3D segmentation result included: (1) cropping or padding the segmentation maps concerning the original size, that is, an inverse operation to the step described in Section 2.3.1; (2) removing anatomically unreasonable artifacts. To remove unreasonable segmentations (e.g., the claustrum does not appear in the first and last slices which contain skull or other tissues), we employed a simple strategy: if there is a claustrum structure detected in the first m and last n ones of a brain along the z-direction, they are considered false positives. Empirically, m and n are set to 20% of the number of axial slices for each scan. The codes and models of the proposed method are made publicly available on *GitHub*.

2.6 | Parameter setting and computation complexity

An appropriate parameter setting is crucial to the successful training of deep convolutional neural networks. We selected the number of epochs to stop the training by contrasting training loss and the performance on validation set over epochs in each experiment, as shown in Figure S2 in the Supplement. Hence, we choose a number of N epochs to avoid over-fitting and to keep a low computational cost by observing the VS and DSC on the validation set. The batch size was empirically set to 30 and the learning rate was set to 0.0002 throughout all experiments by observing the training stability on the validation set.

The experiments are conducted on a GNU/Linux server running Ubuntu 18.04, with 64GB RAM. The number of trainable parameters in the proposed model with one-channel inputs (T1) is 4,641,209. The algorithm was trained on a single NVIDIA Titan-V GPU with 12GB RAM. It takes around 100 min to train a single model for 200 epochs on a training set containing 5000 images with a size of 180×180 pixels. For testing, the segmentation of one scan with 192 slices by an ensemble of two models takes around 90 s using an Intel Xeon CPU (E3-1225v3) (without GPU use). In contrast, the segmentation per scan takes only 3 s when using a GPU.

2.7 | Evaluation metrics and protocol

Three metrics are used to evaluate the segmentation performance in different aspects in the reported experiments. For example, given a ground truth segmentation map G and a predicted segmentation map P generated by an algorithm, the evaluation metrics are defined as follows.

2.7.1 | Volumetric similarity (VS)

Let V_G and V_P be the volumes of region of interests in G and P, respectively. Then the volumetric similarity (VS) in percentage is defined as:

$$VS = 1 - \frac{|V_G - V_P|}{|V_G + V_P|}$$

2.7.2 | Hausdorff distance (95th percentile) (HD95)

$$H(G,P) = \max\left\{\sup_{x \in G}\inf_{y \in P} d(x,y), \sup_{y \in G}\inf_{x \in P} d(x,y)\right\}$$

Where $d(x,y)$ denotes the distance of x and y, sup denotes the *supremum* and *inf* for the infimum. This measures the distance between the two subsets of metric space. It is modified to obtain a robust metric by using the 95th percentile instead of the maximum (100th percentile) distance.

2.7.3 | Dice similarity coefficient

$$DSC = \frac{2(G \cap P)}{|G| + |P|}$$

This measures the overlap between ground truth maps G and prediction maps P.

We use k-fold cross-validation to evaluate the overall performance. In each split, 80% scans from each scanner are pooled into a

training set and the remaining scans as a test set. This procedure is repeated until all of the subjects were used in the testing phase.

3 | RESULTS

3.1 | Manual segmentation: intra-rater variability

In order to set a benchmark accuracy for manual segmentation, intra-rater variability was assessed based on repeated annotations of 20 left and right claustra by the same experienced neuroradiologist. In order to assure independent segmentation, annotations were performed at least 3 months apart. We obtained the intra-rater variability on 20 scans using the metrics VS, DSC, and HD95 and report the following median values with interquartile ranges (IQR): VS: 0.949, [0.928, 0.972]; DSC: 0.667, [0.642, 0.704], HD95: 2.24 mm, [2.0, 2.55]. Notably, the image resolution of all scans is 1.00 mm^3.

3.2 | DL-based segmentation: single-view vs. multi-view

In order to investigate the added value of multi-view information for the proposed system, we compare the segmentation performances of the single-view model (i.e., axial, coronal, or sagittal) with the multi-view ensemble model. To exclude the influence of scanner acquisition, we evaluate our method on the data from one scanner (*Munich-Achieva*), including 103 subjects and perform 5-fold cross-validation for a fair comparison. In each cross-validation split, the single-view CNNs and multi-view CNNs ensemble model are trained on images from the same subjects. Afterwards, they are evaluated on the test cases with respect to the evaluation metrics. Table 2 shows the segmentation performance of each setting. We observed that the sagittal view yields the worst performance among the three views (Figure 4).

We further perform statistical analysis (Wilcoxon signed-rank test) to compare the statistical significance between the proposed single-view CNNs and multi-view CNNs ensemble model. We observed that the two-view (*axial* + *coronal*) approach outperforms single-view ones significantly on HD95 and DSC. We further compared the three-view (axial, coronal, and sagittal) approach with the two-view(axial and coronal) approach and found that they are comparable in terms of VS, and that the two-view approach outperforms the three-view approach in terms of HD95 ($p = .035$) and DSC ($p = .021$). Thus, in the following sections, we use the *axial* + *coronal* two-view segmentation approach to evaluate the method.

3.3 | DL-based segmentation: stratified *k*-fold cross validation

In order to evaluate the general performance of our axial and coronal multi-view technique on the whole dataset, we performed stratified 5-fold cross validation. In each fold, we take 80% of subjects from all scanners, pool them into a training set and use the rest as a test set. Figure 3 and Table 3 show the segmentation performance of three metrics on 181 scans from four scanners, showing its effectiveness with respect to volume measurements and localization accuracy. In order to compare AI-based segmentation performance to the human expert rater benchmark performance, we performed *Wilcoxon signed-rank test* on 20 subjects as mentioned in Section 3.1 with respect to three evaluation metrics (see Table 3). We found no statistical difference between manual and AI-based segmentation with respect to VS, and we observed superior performance of AI-based segmentation with respect to HD95 and Dice score. This result indicates that AI-based segmentation performance is equal or superior to the human expert level.

3.4 | DL-based segmentation: Influence of individual scanners

To evaluate the generalizability of our method to unseen scanners, we present a leave-one-scanner-out study. For the cross-scanner analysis, we use the scanner IDs to split the 181 cases into training and test sets. In each split, the subjects from three scanners are used as a training set while the subjects from the remaining scanner are used as the test set. This procedure is repeated until all the scanners are used as test set. The achieved performance is comparable with the cross-

TABLE 2 Segmentation performances (median values with IQR) of the single-view approaches and multi-view approaches

						p value		
Metrics	**Axial (A)**	**Coronal (C)**	**Sagittal (S)**	**A + C**	**A + C + S**	**A + C vs. A**	**A + C vs. C**	**A + C vs. A + C + S**
VS (%)	94.4 [90.1, 96.7]	94.7 [90.4, 97.3]	79.1 [73.5, 86.4]	93.3 [89.6, 96.9]	92.9 [89.6, 96.5]	.636	**.008**	.231
HD95↓ (mm)	1.73 [1.41, 2.24]	1.41 [1.41, 2.0]	3.21 [2.24, 3.61]	1.41 [1.41, 1.79]	1.73 [1.41, 1.84]	**<.001**	**<.001**	**.035**
DSC (%)	69.7 [66.0, 72.4]	70.0 [67.2, 73.2]	55.2 [45.7, 63.1]	71.8 [68.7, 74.6]	71.0 [68.5, 74.3]	<.001	**<.001**	**.021**

Note: Values in bold denote statistical significance. The combination of axial and coronal views shows its superiority over individual views. Note that we used equal weights for each view in the multi-view ensemble model.
Abbreviations: A, axial; C, coronal; DSC, dice similarity coefficient; HD95, 95th percentile of Hausdorff distance; S, sagittal; VS, volumetric similarity.

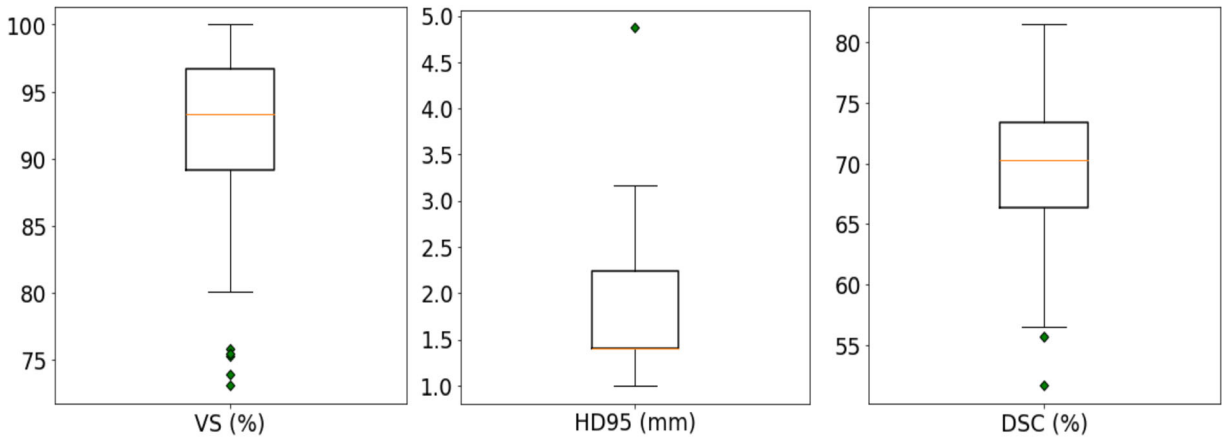


FIGURE 3 Segmentation results of 5-fold cross-validation on the 181 scans across four scanners: *Bonn-Achieva*, *Bonn-Ingenia*, *Munich-Achieva*, and *Munich-Ingenia*. Each box plot summarizes the segmentation performance with respect to one specific evaluation metric

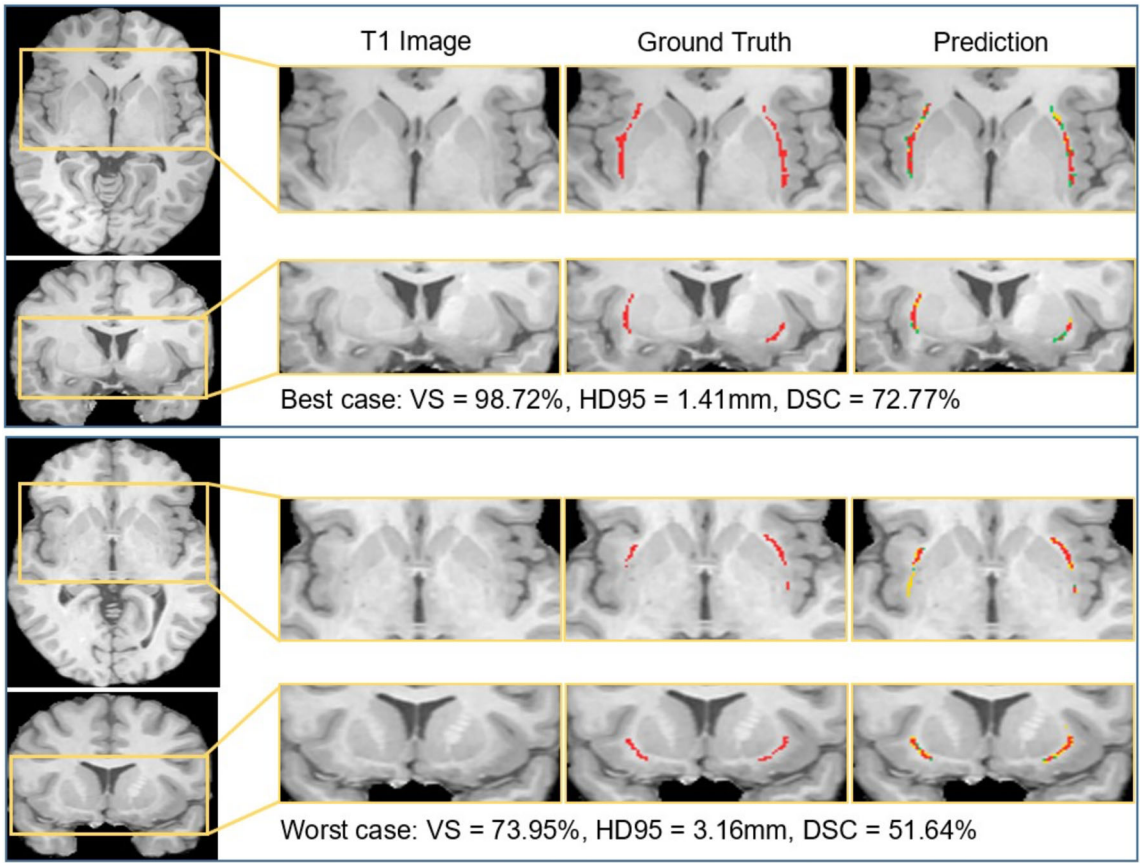


FIGURE 4 Segmentation results of the best case and the worst case in terms of DSC. In the predicted segmentation masks, the red pixels represent true positives, the green ones represent false negatives, and the yellow ones represent false positives

validation results in Section 3.3, where all scanners were seen in the training set. Figure 5 plots the distributions of segmentation performances on four scanners being tested in turns. As shown in Table 4, we found that the cross-validation results achieved significantly lower HD95 and higher DSC than leave-one-scanner-out results at comparable VS. This is because for cross-validation, all scanners are included in training stage and thus no domain shift is seen between training and testing stages. This result indicates that testing the model on unseen scanners hampers segmentation performance.

To further investigate the influence of scanner acquisition for segmentation, we individually perform 5-fold cross-validation on the subsets *Bonn-Ingenia* and *Munich-Achieva* using subject IDs. The other two scanners are not evaluated because they contain relatively fewer scans. We use *Mann–Whitney U test* to compare the performance of

TABLE 3 Performance comparison of manual and AI-based segmentations on 20 subjects with Wilcoxon signed-rank test

Metrics	Manual segmentation [median, IQR]	AI-based segmentation [median, IQR]	*p* value
VS (%)	94.9, [91.4, 97.6]	94.3, [89.6, 96.7]	.821
HD95 (mm)	2.24, [2.0, 2.55]	1.41, [1.41, 2.24]	**.005**
DSC (%)	68.9, [64.2, 70.9]	71.7, [67.8, 73.5]	**.001**

Note: We found that AI-based segmentation performance is equal or superior to the human expert level.
Abbreviations: DSC, Dice similarity coefficient; HD95, 95th percentile of Hausdorff Distance; VS, volumetric similarity.

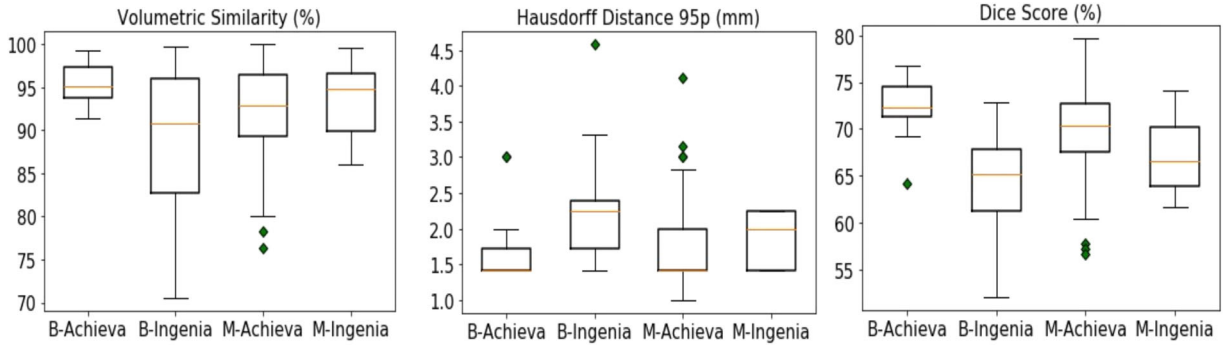


FIGURE 5 Segmentation results of leave-one-scanner-out evaluation on the four scanners. Each sub-figure summarizes the segmentation performance on the testing scans from four scanners with respect to one metric. For example, the boxplot named *Bonn-Achieva* in the left sub-figure shows the distribution of segmentation results on scanner *Bonn-Achieva* (scanner 1) when using data from the other three scanners to train the AI model

TABLE 4 Statistics analysis of leave-one-scanner-out segmentation results and *k*-fold cross-validation results

Metrics	Leave-one-scanner-out (mean ± Std)	*k*-fold cross-validation (mean ± Std)	*p* value
VS (%)	91.9 ± 6.2	92.2 ± 5.7	.268
HD95(mm)↓	1.86 ± 0.58	1.76 ± 0.51	**<.001**
DSC (%)	68.3 ± 5.0	69.5 ± 5.3	**<.001**

Note: Values in bold denote statistical significance. Statistical differences between them with respect to HD95 and Dice score were observed. It indicated that testing on unseen scanners harms the segmentation performance.
Abbreviations: DSC, Dice similarity coefficient; HD95, 95th percentile of Hausdorff Distance; VS, volumetric similarity.

the two groups. We found that *Bonn-Ingenia* obtained significantly lower VS and lower DSC than *Munich-Achieva*, which indicates that scanner characteristics such as image contrast, noise level, etc., generally affect the performance of AI-based segmentation. The box plots of the two evaluations are shown in Figure S1.

3.5 | How much training data is needed?

Since supervised deep learning is a data-driven machine learning method, it commonly requires a large amount of training data to optimize the nonlinear computational model. However, it is necessary to know the boundary when the model begins to saturate in order to avoid unnecessary manual annotations. Here, we perform a quantitative analysis on the effect of the amount of training data. Specifically, we split the 181 scans into a training set and a validation set with a ratio of 4:1 in a stratified manner from four scanners, resulting in 146 subjects for training and 35 for validation. As a start, we randomly pick 10% of the scans from the training set, train, and test the model. Then we gradually increased the size of the training set by a step of 10%. Figure 6 shows that the HD95 and the DSC only marginally improve on the validation set when >50% of the training set is used, while the VS is relatively stable over the whole range. Thus, we conclude that a training set including around 75 manually annotated scans is sufficient to obtain good segmentation results.

4 | DISCUSSION AND CONCLUSION

We have presented a deep-learning-based approach to accurately segment the claustrum, a complex and tiny gray matter structure of the human forebrain that has not been amenable to conventional segmentation methods. The proposed method uses multi-view information from T1-weighted MRI and achieves expert-level segmentation in a fully automated manner. To the best of our knowledge, this is the

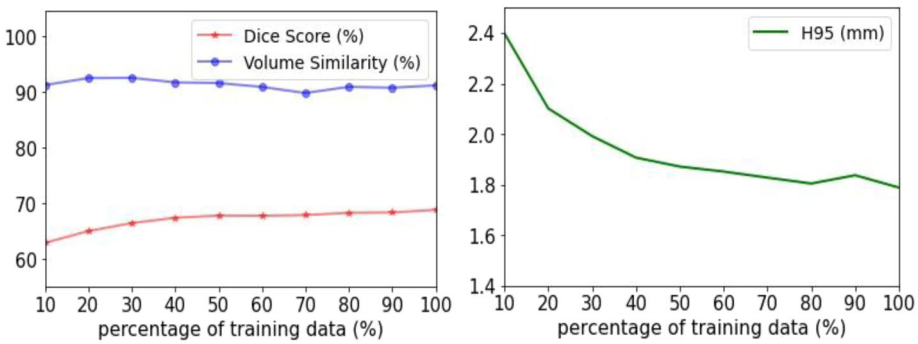


FIGURE 6 Segmentation performance on the validation set when gradually increasing the percentage of the training data by a step of 10%. Only a marginal improvement on the validation set was observed when >50% of the training set was used

first work on fully automated segmentation of human claustrum using state-of-the-art deep learning techniques.

The first finding is that the segmentation performance benefits from leveraging multi-view information, specifically combining axial and coronal orientations. The significance of improvement was confirmed using paired difference tests. The multi-view fusion process imitates the annotation workflow by neuroradiologists, relying on 3D anatomical knowledge from multiple views. This strategy is also shown to be effective in common brain structure segmentation (Wei, Xia, & Zhang, 2019; Zhao, Zhang, Song, & Liu, 2019) and cardiac image segmentation (Mortazi et al., 2017). We observed that integrating sagittal view does not further improve the performance. This is because the claustrum, a thin, sheet-like structure is mainly oriented in the sagittal plane and can hardly be delineated in the sagittal view.

The proposed method yields a high median volumetric similarity, a small Hausdorff distance, and a decent Dice score in the cross-validation experiments. Although the achieved Dice score presents a relatively small value (~70%), we claim that this is excellent considering the structure of the claustrum is very tiny (usually <1500 voxels at 1 mm^3 isotropic resolution). We illustrate the correlation between Dice scores and claustrum volumes in the Supplement. In similar tasks such as segmentation of multiple sclerosis lesions with thousands of voxels, a Dice score of around 75% would be considered excellent. For the segmentation of larger tissues such as white matter and gray matter, Dice scores would reach 95% (Gabr et al., 2020). Nevertheless, HD95 quantifies the distance between prediction and ground-truth masks and is robust to assess tiny and thin structures (Kuijf et al., 2019).

Another valuable finding is that the proposed algorithm achieves expert-level segmentation performance and even outperforms a human expert rater in terms of DSC and HD95, which is confirmed by comparing the two groups of segmentation performances done by human rater and the proposed method. We conclude that the human rater presents more bias when the structure is tiny and ambiguous. Meanwhile, an AI-based algorithm learns to fit the available knowledge and shows a stable behavior when performing automated segmentation. This finding aligns recent advances in biomedical research where deep learning-based methods demonstrate unbiased quantification of structures (Todorov et al., 2020). Thus, we conclude that the proposed method would quantify the claustrum structure in an accurate and unbiased way.

We found that the segmentation performance slightly dropped when the AI-based model was tested on unseen scanners. Domain shift is commonly observed in machine learning tasks between training and testing data with different distributions. However, from our observation, the performance drop in the experiment is not severe, and the segmentation outcome is satisfactory. This is because scanners are in similar resolution from the same manufacturer, and the scans are properly pre-processed, resulting in a small domain gap. To enforce our model to be generalized to unseen scanners, domain adaptation methods (Dou, de Castro, Kamnitsas, & Glocker, 2019; Kamnitsas et al., 2017) are to be investigated in future studies.

Although the proposed method reaches expert-level performance and provides unbiased quantification results, our work has a few limitations. First, the human claustrum has a thin and sheet-like structure. Thus, high-resolution imaging as used in this study at an isotropic resolution of 1 mm^3 will result in partial volume effects, which significantly affect both the manual expert annotation and the automated segmentation. We addressed this bias by using a clear segmentation protocol to reduce variability in manual annotations as the reference standard. Second, the data distribution of the four datasets is highly imbalanced. It potentially affects the accuracy of the leave-one-scanner-out experiment in Section 3.4, especially when a significant sub-set (e.g., Munich-2) was taken out as a test set. In future work, evaluating the scanner influence on a more balanced dataset would avoid such an effect.

In conclusion, we described a multi-view deep learning approach for automatic segmentation of human claustrum structure. We empirically studied the effectiveness of multi-view information, leave-one-scanner-out study, the influence of imaging protocols and the effect of the amount of training data. We found that: (1) multi-view information, including coronal and axial views, provide complementary information to identify the claustrum structure; (2) multi-view automatic segmentation is equal or superior to manual segmentation accuracy; (3) scanner type affects segmentation accuracy even for identical sequence parameter settings; (4) a training set with 75 scans and annotation is sufficient to achieve satisfactory segmentation result. We have made our Python implementation codes available on *GitHub*.

ACKNOWLEDGMENTS

We thank all current and former members of the Bavarian Longitudinal Study Group who contributed to general study organization, recruitment and data collection, management and subsequent

analyses, including (in alphabetical order): Barbara Busch, Stephan Czeschka, Claudia Grünzinger, Christian Koch, Diana Kurze, Sonja Perk, Andrea Schreier, Antje Strasser, Julia Trummer, and Eva van Rossum. We are grateful to the staff of the Department of Neuroradiology in Munich and the Department of Radiology in Bonn for their help in data collection. Most importantly, we thank all our study participants and their families for their efforts to take part in this study. This study is supported by the Deutsche Forschungsgemeinschaft (SO 1336/1-1 to Christian Sorg), German Federal Ministry of Education and Science (BMBF 01ER0803 to Christian Sorg) and the Kommission für Klinische Forschung, Technische Universität München (KKF 8765162 to Christian Sorg). Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Hongwei Li https://orcid.org/0000-0002-5328-6407
Aurore Menegaux https://orcid.org/0000-0003-3965-0396
Benita Schmitz-Koep https://orcid.org/0000-0002-8874-5749
Dennis Hedderich https://orcid.org/0000-0001-8994-5593

REFERENCES

Aljabar, P., Heckemann, R. A., Hammers, A., Hajnal, J. V., & Rueckert, D. (2009). Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy. *NeuroImage*, *46*(3), 726–738.

Aljabar, P., Wolz, R., & Rueckert, D. (2012). Manifold learning for medical image registration, segmentation, and classification. *Machine Learning in Computer-Aided Diagnosis: Medical Imaging Intelligence and Analysis*, *1*, 351–372.

Arrigo, A., Mormina, E., Calamuneri, A., Gaeta, M., Granata, F., Marino, S., … Quartarone, A. (2017). Inter-hemispheric claustral connections in human brain: A constrained spherical deconvolution-based study. *Clinical Neuroradiology*, *27*(3), 275–281.

Berman, S., Schurr, R., Atlan, G., Citri, A., & Mezer, A. A. (2020). Automatic segmentation of the dorsal claustrum in humans using in vivo high-resolution MRI. *Cerebral Cortex Communications*, *1*(1), 1–14. https://doi.org/10.1093/texcom/tgaa062

Brown, S. P., Mathur, B. N., Olsen, S. R., Luppi, P.-H., Bickford, M. E., & Citri, A. (2017). New breakthroughs in understanding the role of functional interactions between the neocortex and the claustrum. *Journal of Neuroscience*, *37*(45), 10877–10881.

Bruguier, H., Suarez, R., Manger, P., Hoerder-Suabedissen, A., Shelton, A. M., Oliver, D. K., … Puelles, L. (2020). In search of common developmental and evolutionary origin of the claustrum and subplate. *Journal of Comparative Neurology*, *528*(17), 2956–2977.

Chen, H., Dou, Q., Yu, L., Qin, J., & Heng, P.-A. (2018). VoxResNet: Deep voxelwise residual networks for brain segmentation from 3D MR images. *NeuroImage*, *170*, 446–455.

Crick, F. C., & Koch, C. (2005). What is the function of the claustrum? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1458), 1271–1279.

Davis, W. B. (2008). The claustrum in autism and typically developing male children: a quantitative MRI study. Brigham Young University.

Dou, Q., de Castro, D. C., Kamnitsas, K. & Glocker, B. (2019). *Domain generalization via model-agnostic learning of semantic features*. Paper presented at the Advances in Neural Information Processing Systems 32, Vancouver, Canada, 6450-6461.

Gabr, R. E., Coronado, I., Robinson, M., Sujit, S. J., Datta, S., Sun, X., … Narayana, P. A. (2020). Brain and lesion segmentation in multiple sclerosis using fully convolutional neural networks: A large-scale study. *Multiple Sclerosis Journal*, *26*(10), 1217–1226.

Goll, Y., Atlan, G., & Citri, A. (2015). Attention: The claustrum. *Trends in Neurosciences*, *38*(8), 486–495.

Hedderich, D. M., Bäuml, J. G., Berndt, M. T., Menegaux, A., Scheef, L., Daamen, M., … Wolke, D. (2019). Aberrant gyrification contributes to the link between gestational age and adult IQ after premature birth. *Brain*, *142*(5), 1255–1269.

Heimann, T., & Meinzer, H.-P. (2009). Statistical shape models for 3D medical image segmentation: A review. *Medical Image Analysis*, *13*(4), 543–563.

Hoerder-Suabedissen, A., & Molnár, Z. (2015). Development, evolution and pathology of neocortical subplate neurons. *Nature Reviews Neuroscience*, *16*(3), 133–146.

Iglesias, J. E., Liu, C.-Y., Thompson, P. M., & Tu, Z. (2011). Robust brain extraction across datasets and comparison with publicly available methods. *IEEE Transactions on Medical Imaging*, *30*(9), 1617–1634.

Johnson, J. I., & Fenske, B. A. (2014). History of the study and nomenclature of the claustrum. In *The Claustrum* (pp. 1-27). Amsterdam: Elsevier.

Joshi, A. A., Shattuck, D. W., Thompson, P. M., & Leahy, R. M. (2007). Surface-constrained volumetric brain registration using harmonic mappings. *IEEE Transactions on Medical Imaging*, *26*(12), 1657–1669.

Kamnitsas, K., Baumgartner, C., Ledig, C., Newcombe, V., Simpson, J., Kane, A., Rueckert, D. & Glocker, B. (2017). *Unsupervised domain adaptation in brain lesion segmentation with adversarial networks*. Paper presented at the International Conference on Information Processing in Medical Imaging (pp. 597-609). Springer, Cham.

Kamnitsas, K., Ledig, C., Newcombe, V. F., Simpson, J. P., Kane, A. D., Menon, D. K., … Glocker, B. (2017). Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Medical Image Analysis*, *36*, 61–78.

Kowiański, P., Dziewiątkowski, J., Kowiańska, J., & Moryś, J. (1999). Comparative anatomy of the claustrum in selected species: A morphometric analysis. *Brain, Behavior and Evolution*, *53*(1), 44–54.

Krimmel, S. R., White, M. G., Panicker, M. H., Barrett, F. S., Mathur, B. N., & Seminowicz, D. A. (2019). Resting state functional connectivity and cognitive task-related activation of the human claustrum. *NeuroImage*, *196*, 59–67.

Kuijf, H. J., Biesbroek, J. M., De Bresser, J., Heinen, R., Andermatt, S., Bento, M., … Casamitjana, A. (2019). Standardized assessment of automatic segmentation of white matter hyperintensities and results of the WMH segmentation challenge. *IEEE Transactions on Medical Imaging*, *38*(11), 2556–2568.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Li, H., Jiang, G., Zhang, J., Wang, R., Wang, Z., Zheng, W.-S., & Menze, B. (2018). Fully convolutional network ensembles for white matter hyperintensities segmentation in MR images. *NeuroImage*, *183*, 650–665.

Manjón, J. V., Coupé, P., Martí-Bonmatí, L., Collins, D. L., & Robles, M. (2010). Adaptive non-local means denoising of MR images with spatially varying noise levels. *Journal of Magnetic Resonance Imaging*, *31*(1), 192–203.

Mathur, B. N. (2014). The claustrum in review. *Frontiers in Systems Neuroscience*, *8*, 48.

Milardi, D., Bramanti, P., Milazzo, C., Finocchio, G., Arrigo, A., Santoro, G., … Gaeta, M. (2015). Cortical and subcortical connections of the human claustrum revealed in vivo by constrained spherical deconvolution tractography. *Cerebral Cortex*, *25*(2), 406–414.

Milletari, F., Navab, N. & Ahmadi, S.-A. (2016). *V-net: Fully convolutional neural networks for volumetric medical image segmentation*. Paper presented at the 2016 Fourth International Conference on 3D vision (3DV) (pp. 565-571). Piscataway, NJ:IEEE.

Mortazi, A., Karim, R., Rhode, K., Burt, J., & Bagci, U. CardiacNET: *Segmentation of left atrium and proximal pulmonary veins from MRI using multi-view CNN*. Paper presented at the International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham,Switzerland: Springer.

Pearson, R., Brodal, P., Gatter, K., & Powell, T. (1982). The organization of the connections between the cortex and the claustrum in the monkey. *Brain Research*, *234*(2), 435–441.

Prados, F., Ashburner, J., Blaiotta, C., Brosch, T., Carballido-Gamio, J., Cardoso, M. J., … De Leener, B. (2017). Spinal cord grey matter segmentation challenge. *NeuroImage*, *152*, 312–329.

Puelles, L. (2014). Development and evolution of the claustrum. In *The Claustrum* (pp. 119–176). Amsterdam, The Netherlands: Elsevier.

Remedios, R., Logothetis, N. K., & Kayser, C. (2010). Unimodal responses prevail within the multisensory claustrum. *Journal of Neuroscience*, *30*(39), 12902–12907.

Remedios, R., Logothetis, N. K., & Kayser, C. (2014). A role of the claustrum in auditory scene analysis by reflecting sensory change. *Frontiers in Systems Neuroscience*, *8*, 44.

Reser, D. H., Majka, P., Snell, S., Chan, J. M., Watkins, K., Worthy, K., … Rosa, M. G. (2017). Topography of claustrum and insula projections to medial prefrontal and anterior cingulate cortices of the common marmoset (*Callithrix jacchus*). *Journal of Comparative Neurology*, *525*(6), 1421–1441.

Ronneberger, O., Fischer, P., & Brox, T. (2015). *U-net: Convolutional networks for biomedical image segmentation*. Paper presented at the International Conference on Medical Image Computing and Computer-Assisted Intervention. (pp. 234–241). Cham, Switzerland: Springer.

Todorov, M. I., Paetzold, J. C., Schoppe, O., Tetteh, G., Efremov, V., Völgyi, K., … Menze, B. (2020). Automated analysis of whole brain vasculature using machine learning. *Nature Methods*, *17*(4), 442–449.

Torgerson, C. M., Irimia, A., Goh, S. M., & Van Horn, J. D. (2015). The DTI connectivity of the human claustrum. *Human Brain Mapping*, *36*(3), 827–838.

Torgerson, C. M., & Van Horn, J. D. (2014). A case study in connectomics: The history, mapping, and connectivity of the claustrum. *Frontiers in Neuroinformatics*, *8*, 83.

Wachinger, C., Reuter, M., & Klein, T. (2018). DeepNAT: Deep convolutional neural network for segmenting neuroanatomy. *NeuroImage*, *170*, 434–445.

Watson, C., & Puelles, L. (2017). Developmental gene expression in the mouse clarifies the organization of the claustrum and related endopiriform nuclei. *Journal of Comparative Neurology*, *525*(6), 1499–1508.

Wei, J., Xia, Y., & Zhang, Y. (2019). M3Net: A multi-model, multi-size, and multi-view deep neural network for brain magnetic resonance image segmentation. *Pattern Recognition*, *91*, 366–378.

Yushkevich, P. A., Piven, J., Hazlett, H. C., Smith, R. G., Ho, S., Gee, J. C., & Gerig, G. (2006). User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *NeuroImage*, *31*(3), 1116–1128.

Zhao, Y.-X., Zhang, Y.-M., Song, M. & Liu, C.-L. (2019). *Multi-view semi-supervised 3d whole brain segmentation with a self-ensemble network*. Paper presented at the International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 256-265). Springer, Cham.

Zingg, B., Dong, H. W., Tao, H. W., & Zhang, L. I. (2018). Input–output organization of the mouse claustrum. *Journal of Comparative Neurology*, *526*(15), 2428–2443.

Zingg, B., Hintiryan, H., Gou, L., Song, M. Y., Bay, M., Bienkowski, M. S., … Toga, A. W. (2014). Neural networks of the mouse neocortex. *Cell*, *156*(5), 1096–1111.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Li, H., Menegaux, A., Schmitz-Koep, B., Neubauer, A., Bäuerlein, F. J. B., Shit, S., Sorg, C., Menze, B., & Hedderich, D. (2021). Automated claustrum segmentation in human brain MRI using deep learning. *Human Brain Mapping*, *42*(18), 5862–5872. https://doi.org/10.1002/hbm.25655

4 Efficient neonate claustrum segmentation with deep transfer learning

This chapter has been published as a **peer-reviewed journal publication**:

[1] A. Neubauer*, **H. Li***, J. Wendt, B. Schmitz-Koep, A. Menegaux, D. Schinz, B. Menze, C. Zimmer, C. Sorg, and D. M. Hedderich. "Efficient claustrum segmentation in T2-weighted neonatal brain MRI using transfer learning from adult scans". In: ***Clinical Neuroradiology*** (2022), pp. 1–12

Synopsis: This work explores efficient deep learning technique to reduce manual annotation effort for image segmentation tasks. It establishes a transfer learning based framework for claustrum segmentation. Specifically, the learned knowledge from *adult* claustrum segmentation in *T1-w* MR scans (one work in Chapter 3) is transferred to *neonate* claustrum segmentation in *T2-w* ones. The effectiveness of the transfer learning technique was demonstrated in comparison with nontransfer learning. The model can achieve satisfactory segmentation with only 12 annotated scans. Finally, the model's applicability was verified on 528 scans and revealed reliable segmentations in 97.4%. The developed fast and accurate automated segmentation has great potential in large-scale study cohorts and to facilitate MRI-based connectome research of the neonatal claustrum.

Contributions of thesis author: algorithm design and implementation, computational experiments and composition of manuscript.

Clin Neuroradiol
https://doi.org/10.1007/s00062-021-01137-8

ORIGINAL ARTICLE

Efficient Claustrum Segmentation in T2-weighted Neonatal Brain MRI Using Transfer Learning from Adult Scans

Antonia Neubauer[1,2] · Hongwei Bran Li[3,4] · Jil Wendt[1,2] · Benita Schmitz-Koep[1,2] · Aurore Menegaux[1,2] · David Schinz[1,2] · Bjoern Menze[3,4] · Claus Zimmer[1,2] · Christian Sorg[1,2,5] · Dennis M. Hedderich[1,2]

Received: 11 October 2021 / Accepted: 25 December 2021

Abstract

Purpose Intrauterine claustrum and subplate neuron development have been suggested to overlap. As premature birth typically impairs subplate neuron development, neonatal claustrum might indicate a specific prematurity impact; however, claustrum identification usually relies on expert knowledge due to its intricate structure. We established automated claustrum segmentation in newborns.

Methods We applied a deep learning-based algorithm for segmenting the claustrum in 558 T2-weighted neonatal brain MRI of the developing Human Connectome Project (dHCP) with transfer learning from claustrum segmentation in T1-weighted scans of adults. The model was trained and evaluated on 30 manual bilateral claustrum annotations in neonates.

Results With only 20 annotated scans, the model yielded median volumetric similarity, robust Hausdorff distance and Dice score of 95.9%, 1.12 mm and 80.0%, respectively, representing an excellent agreement between the automatic and manual segmentations. In comparison with interrater reliability, the model achieved significantly superior volumetric similarity ($p = 0.047$) and Dice score ($p < 0.005$) indicating stable high-quality performance. Furthermore, the effectiveness of the transfer learning technique was demonstrated in comparison with nontransfer learning. The model can achieve satisfactory segmentation with only 12 annotated scans. Finally, the model's applicability was verified on 528 scans and revealed reliable segmentations in 97.4%.

Conclusion The developed fast and accurate automated segmentation has great potential in large-scale study cohorts and to facilitate MRI-based connectome research of the neonatal claustrum. The easy to use models and codes are made publicly available.

Keywords Claustrum · Newborn infants · Deep learning · Image segmentation · Transfer learning

The authors Antonia Neubauer and Hongwei Bran Li contributed equally to the manuscript.

Data Availability The data that support the findings of this study are available on the dHCP website (http://www.developingconnectome.org/project/). The models and codes are made publicly available (https://github.com/hongweilibran/claustrum_multi_view).

✉ Antonia Neubauer
neu-antonia@web.de

[1] Department of Diagnostic and Interventional Neuroradiology, Klinikum rechts der Isar, Technical University of Munich, Ismaninger Strasse 22, 81675 Munich, Germany

[2] TUM-NIC Neuroimaging Center, Munich, Germany

[3] Department of Informatics, Technical University of Munich, Munich, Germany

[4] Department of Quantitative Biomedicine, University of Zurich, Zurich, Switzerland

[5] Department of Psychiatry, Klinikum rechts der Isar, Technical University of Munich, Munich, Germany

Published online: 24 January 2022

Abbreviations

AS	Automated segmentation
CPU	Central processing unit
DA	Data augmentation
dHCP	Developing Human Connectome Project
DSC	Dice similarity coefficient
GA	Gestational age
GPU	Graphics processing unit
HD95	95th percentile of the Hausdorff distance
IQR	Interquartile range
MRI	Magnetic resonance imaging
non-TL	Nontransfer learning
T2-w	T2-weighted
TL	Transfer learning
VS	Volumetric similarity

Introduction

The claustrum is a thin and sheet-like gray matter structure of the mammalian forebrain between the striatum and insular cortex, or more precisely, in humans between the external and extreme capsule [1, 2]. Examining the claustrum is challenging due to its small size, ambiguous shape, and deep brain location. The function of the claustrum remains unclear, and most investigations are based on animal studies, which highlights the need for imaging-based studies in humans. Preliminary findings suggest that the claustrum is relevant for consciousness [3], task switching, salience network organization, attention guiding, and top-down control [4–8]. Human studies suggest a role of the claustrum in selective attention and task switching [9]; however, these investigations are usually limited to small sample sizes [10, 11]. In large cohorts, common manual claustrum segmentation would be very laborious and time consuming.

Moreover, there is a lack of knowledge about claustrum development in humans. Most studies focus on animals, while macrostructural and microstructural maturation in humans remain unknown [1, 12, 13]. It has been shown that there are significant differences between very preterm and term-born young adults in patterns of BOLD activity in clusters centered on the claustrum during a learning task [14]. A clear rationale to study claustrum development, particularly in premature-born neonates, comes from its shared ontogenetic trajectory with so-called subplate neurons [15]. The subplate neurons are a predominantly transient cell population and are therefore vulnerable to hypoxic-ischemic events and thus, play a key pathophysiologic role for disturbed neurodevelopment after premature birth [16–21]. This is underlined by a previous study showing altered claustrum microstructure in premature-born adults [22], which is a finding with potentially significant implications. Examination of the claustrum and altered claustrum structure in neurodevelopmental disorders such as impaired development after premature birth may lead to the establishment of imaging biomarkers for subplate neuron pathology. This may also be extended to other neurodevelopmental disorders with presumed subplate neuron pathology, such as schizophrenia and autism spectrum disorder [23]. Hence, close examination and characterization of claustrum development in younger cohorts is of special interest; however, data about the claustrum in a sizable neonatal cohort are missing, mostly due to the lack of adequate automated segmentation methods.

Recently, automated segmentation of the human claustrum in adults has been investigated by structural approximation to the dorsal claustrum [24] and a two-dimensional deep-learning approach [25]. Furthermore, a multiview deep learning-based model has been proposed [26] to segment the human claustrum trained on a large annotated dataset; however, no reliable automated segmentation method for the claustrum in neonatal MRI exists.

To fill this gap, this study presents an efficient deep learning-based segmentation framework using manual expert annotations of the claustrum in a sophisticated cohort of neonatal MRI from the developing Human Connectome Project (dHCP) [27] comprising ongoing brain development. Transfer learning [28] enabled reuse of available artificial intelligence models despite different neuroanatomy, scanner, image sequence, and image resolution shift, and drastically shortened the training time to 90 min. Segmentation accuracy was evaluated based on three canonical performance metrics, volumetric similarity (VS), 95th percentile of the Hausdorff distance (HD95), and Dice similarity coefficient (DSC), and compared with intrarater and interrater reliability of manual segmentation. The proposed technique was also compared to a nontransfer learning approach. The study provides an insight into the training process by quantifying the amount of manually annotated images needed for good segmentation results. Lastly, the deep learning model was applied to the whole, large-scale dHCP dataset to see how its output holds out against rigorous visual quality control. An accuracy drop in young neonates was analyzed and solved by an age-stratified training set. Training and testing code and models are released on GitHub for other research groups. A detailed claustrum segmentation protocol is in the Online Supplement. In parallel, the proposed transfer learning approach serves as a template for similar segmentation tasks of intricate and small structures in the developing brain.

Material and Methods

In the following parts, the single term “model” refers to a 2D artificial neural network while “combined model” in-

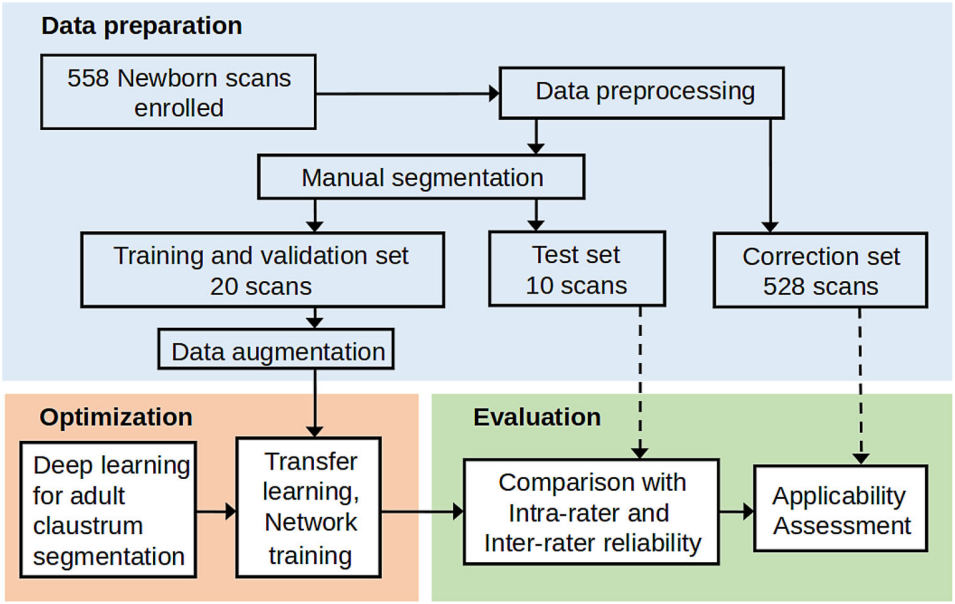


Fig. 1 A schematic view of the image segmentation and evaluation pipeline of this study. It includes three stages: 1) data preparation, 2) model optimization and 3) framework evaluation

tegrates several 2D models (*see* Section Multiview Convolutional Neural Network). Whereas manually acquired tracing of the claustrum is always described with the term "manual segmentation", the output of a model is described as "automated segmentation" or "prediction" in an interchangeable way.

The general image processing diagram in this work includes three stages shown in Fig. 1. Data preparation deals with the enrolment of 558 subjects, image preprocessing and manual segmentation of neonate claustrum. Optimization aims to perform transfer learning and train a deep-learning model with the manual segmentations provided in the first stage. Finally, the evaluation investigates the effectiveness and the applicability of the established model on unseen data including failure analysis and model improvement. The following two sections describe the datasets and evaluation metrics in this study.

Datasets

All 558 three-dimensional MRI scans of newborns from the second data release of the developing Human Connectome Project (dHCP)[1] were included. The large-scale public dataset contains 558 brain MRI of 505 neonates from 23 to 44 weeks postconceptional age with a mean (±standard deviation) scan age of 40 (±3) gestational weeks. In detail, the study comprises 378 scans of term-born neonates and 180 scans of preterm-born neonates, including 82 scans of very preterm-born neonates (birth age <32 gestational weeks). Data involve previously known at risk groups for neurodevelopmental disorders and incidental findings in clinically unsuspicious neonates [29, 30]. The explicit inclusion and exclusion criteria are shown on the dHCP website[2]. Recruitment and scanning took place at the Evelina Newborn Imaging Centre, St Thomas' Hospital in London, UK [29]. Written consent by the parents was previously requested [27]. Due to immature structures with different tissue composition than in adults, the preferred structural image sequence in neonatal brain MRI are T2-weighted (T2-w) scans. Thus, the dHCP favored this sequence in data preprocessing steps [29] and we focused on it for our study. Images were acquired with a 3T Philips Achieva with a repetition time TR = 12,000 ms and echo time TE = 156 ms, isotropic reconstructed voxel size of 0.5 mm and scanning in axial (SENSE factor: 2.11) and sagittal (SENSE factor: 2.60) plane with a neonatal 32 channel head coil [27]. The structural brain images passed visual quality control, brain extraction, and were preprocessed by retrospective motion and bias correction by the dHCP [29, 31].

Out of this dataset, 30 randomly chosen subjects passed manual segmentation. Subsequently, these scans were split in a training set of 20 subjects and a test set comprising 10 scans for evaluation. The remaining 528 scans served as correction set and did not undergo manual segmentation. Training, test, and correction sets are consistent throughout the experiments (Table 1).

The manual segmentation was performed with ITK-SNAP-v3.6.0[3] [32] on a Wacom Intuos M tablet (Wacom,

[1] http://www.developingconnectome.org/.

[2] http://www.developingconnectome.org/study-inclusion-and-exclusion-criteria/.

[3] http://www.itksnap.org.

Table 1 Characteristics of the dataset in this study. The dataset consists of 558 subjects from the developing Human Connectome Project. For each dataset, the count of scans and the mean scan age (range) in gestational weeks are given

Scanner	Field strength	Voxel size (mm^3)	Training set; scan age	Test set; scan age	Correction set; scan age
Philips Achieva (Philips, Best, The Netherlands)	3T	$0.5 \times 0.5 \times 0.5$	20 scans 39.9 (36.1–42.6)	10 scans 40.4 (38.7–42.3)	528 scans 40.0 (29.3–45.1)

Kazo, Saitama, Japan). The first rater was under close supervision of a board-certified neuroradiologist with 10 years of experience including imaging for a neonatal intensive care unit and 5 years of experience pertaining to imaging of premature-born individuals and related outcomes. The detailed segmentation protocol, which assures a constant structure for more objective and stable results, is described in the Online Supplement. Despite this approach, it remains challenging to define the exact boundaries of the small claustrum due to the ambiguity. To quantify the intrarater reliability of manual segmentation, the first rater traced the right and left claustrum of the 10 subjects in the test set at two time points. Furthermore, these 10 subjects were manually segmented by a second rater with the same protocol to assess interrater reliability.

Model Evaluation

Given a manual segmentation mask M and a predicted segmentation mask P, three different evaluation metrics assessed the model performance:

Volumetric Similarity (VS)

While V_M and V_P are the volumes of the claustrum in M and P, respectively, the volumetric similarity (VS) between them is defined as:

$$VS[\%] = 1 - \frac{|V_M - V_P|}{|V_M + V_P|}$$

95th Percentile of the Hausdorff Distance (HD95)

The Hausdorff distance (HD) is a common score to measure the surface distance between two masks M and P [33]:

$$HD\,(M, P) = \max\{\sup_{x \in M} \inf_{y \in P} d\,(x, y)\,,\ \sup_{y \in P} \inf_{x \in M} d\,(x, y)\}$$

$d(x,y)$ denotes the Euclidean distance of x and y, *sup* terms the supremum and *inf* the infimum. We used the 95th percentile instead of the maximum (100th percentile) distance to discount single outliers.

Dice Similarity Coefficient (DSC)

$$DSC = \frac{2\,(M \cap P)}{|M| + |P|}$$

The Dice similarity coefficient (DSC) quantifies the spatial overlap between manual segmentation M and prediction mask P.

Evaluation Protocol

K-fold Cross-validation The model's overall performance was evaluated with k-fold cross-validation with 20 subjects in the training/validation set. While k was set to 5, in each split 80% of the scans were pooled into the training set and the remaining 20% were used for validation. After five iterations, all subjects were evaluated in the validation phase.

Evaluation on a Test Set The model was optimized on 20 subjects. The combined model was evaluated on a test set with 10 subjects and compared with intrarater and interrater reliability.

Applicability Assessment The combined model was applied to the correction set with 528 subjects. These predictions were compared with their subsequently manually corrected correlates.

Additional Preprocessing and Postprocessing

Image Preprocessing We performed additional steps on top of the basic preprocessing steps carried out by dHCP protocol (Sect. Datasets). First, a z-score normalization standardized the brain voxel intensities for each scan as proposed in [26]. Second, every slice was cropped to a uniform size of 200×200 pixels to exclude background information. Third, the first and last 25% of the slices were removed based on empirical decision to focus on central parts of the brain, which include the claustrum, and to lower the computational time.

Segmentation Postprocessing After generating a segmentation, two postprocessing steps were applied to it: 1) the segmentation maps were padded with respect to the original size, i.e., an inverse operation to the previous second

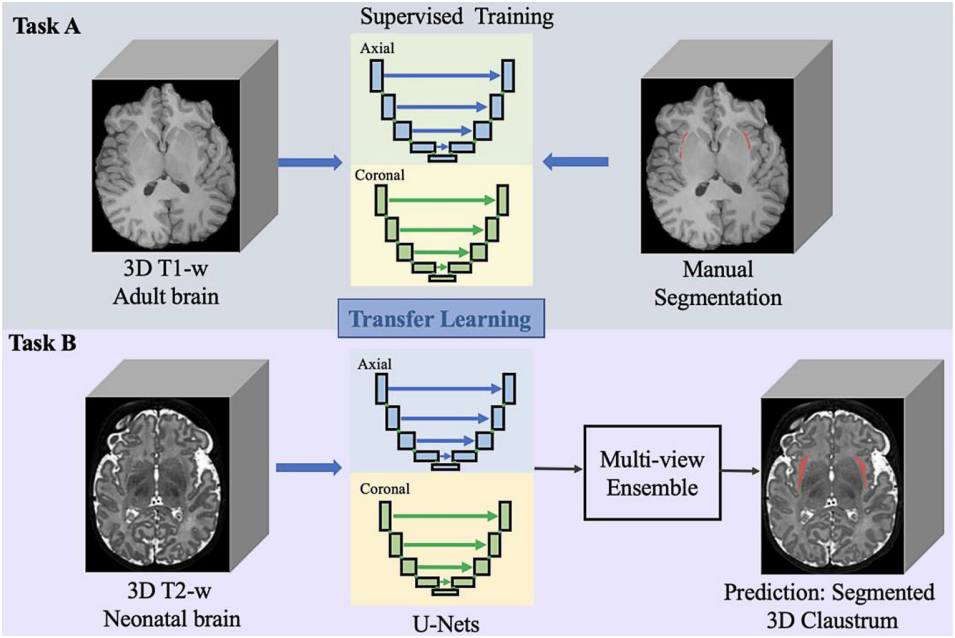


Fig. 2 A schematic view of the proposed segmentation method using transfer learning and multiview convolutional neural networks to segment the newborn claustrum given limited data. The network for each view (i.e., axial and coronal) is a 2D convolutional network architecture, and it takes the raw images as the input and predicts the claustrum segmentation

preprocessing step and 2) an according sequence to preprocessing step three to remove some artifacts.

Data Augmentation

In contrast to expensive manual segmentation, data augmentation (DA) is a method to enlarge the amount and the diversity of training data. A stack of selective transformations, including moderate shift, scaling, rotation, and shearing to the image slices and the corresponding masks, resulted in doubled training data (*see* Fig. S1 in the Online Supplement for selection of DA methods). For comparison, the same models were trained with and without DA and their performance was assessed on the validation set. There was no significant difference regarding the VS; however, DA led to a significant improvement of automated segmentation concerning HD95 and DSC (see Table S1). For the stated reasons, data augmentation enriched the following experiments.

Multiview Convolutional Neural Network

As automated neonatal claustrum segmentation is not feasible to conventional atlas-based methods, we adopted a supervised deep-learning approach developed for adults [26]. While training, the model takes labeled slices of MR images as input data and adapts its parameters towards accurate prediction by minimizing the loss function (Sect. Parameter Setting and Computation Complexity). Finally, the trained model can be applied to trace the claustrum in unseen neonatal images. Based on the beneficial multiview approach proposed in [26], we train coronal and axial deep convolutional neural networks on 2D single-view slices after parsing 3D MRI volume into axial and coronal views. In the test stage, the predictions are automatically combined on a voxel-wise level.

The network architecture of the convolutional neural network [26] adapted to the neonatal image format is shown in Fig. S2. It has a U-shape [34] with a down-convolutional part that extracts features of the T2-w input scans. The up-convolutional part assigns the categories claustrum or non-claustrum to each pixel conforming a segmentation of the claustrum.

Transfer Learning

Transfer learning is typically performed using a designed model and pretrained weights from one source task and fine-tuning on the target task. In this work, the knowledge from task A: human claustrum segmentation in T1-w adult images, was transferred to task B**:** claustrum segmentation in high-resolution T2-w images of neonates scanned in a range of 21 gestational weeks with ongoing brain development. As shown in Fig. 2, we used the same model and directly took its weights learned from task A. Then the multiview networks were optimized with only 20 T2-w scans with manual segmentations for task B. It took around 90 min for the whole training process and 6s for automated segmentation using a common NVIDIA (Santa Clara, CA, USA) graphics processing unit (GPU). The high efficiency of our framework is explained in the following sections.

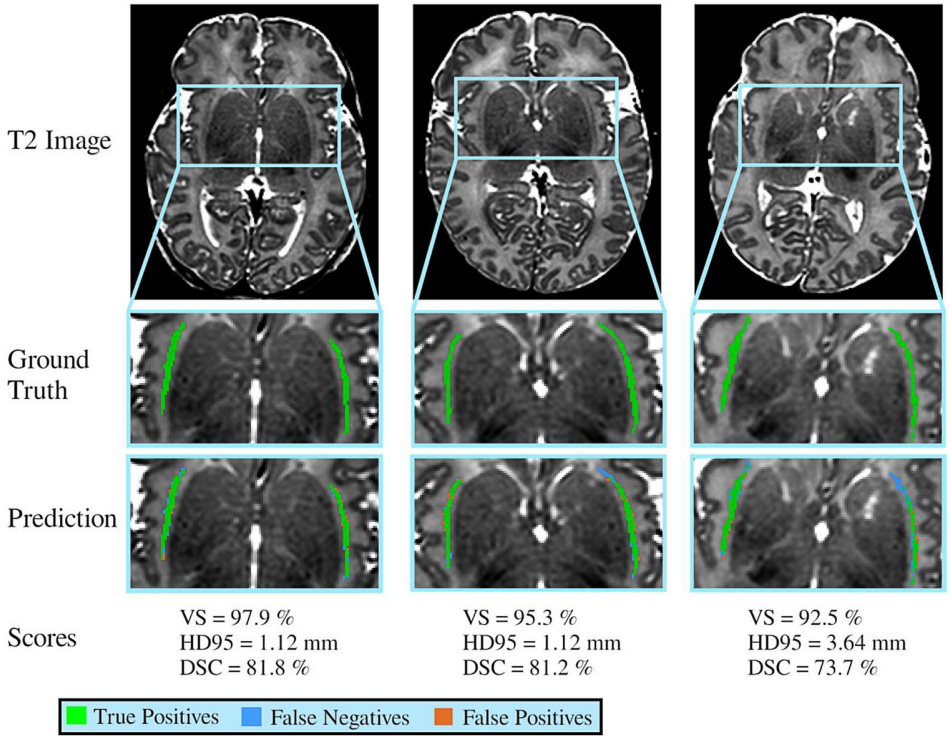


Fig. 3 Segmentation results of three sample cases. In the automated segmentation masks, the green pixels represent true positives, the blue ones represent false negatives, and orange ones represent false positives. Examples are sorted according to accuracy as determined by the Dice similarity coefficient (DSC). *VS* volumetric similarity, *HD95* 95th percentile of Hausdorff distance

Parameter Setting and Computation Complexity

The hyperparameters were chosen consistently for all experiments and optimized efficiency and accuracy. Each model was trained for 30 epochs to avoid overfitting and to keep a low computational cost by monitoring VS and DSC on a validation set. The batch size was empirically set to 60 as a relatively large batch size tended to a more stable training than a smaller batch size mainly due to the imbalanced nature of the training set. The learning rate was set to 0.0002. Non-TL models, which were prepared for comparison reasons, were trained for 275 epochs (see Fig. S5). The other hyperparameters were similar as for TL.

In the claustrum segmentation task, the distribution of claustrum voxels and non-claustrum voxels are highly imbalanced. To handle this issue, the Dice loss was used as a loss function to minimize the difference between manual segmentation and prediction during training [26, 35, 36].

All experiments were performed on a Linux workstation running Ubuntu 20.04 (Canonical Ltd., London, UK), with 64 GB RAM. The number of trainable parameters in the single-view architecture is 2,494,529. The model was trained on one NVIDIA Titan-Xp GPU with 12 GB of GDDR5X memory. Training a single model for 30 epochs on a training set containing 4200 images with a size of 200 × 200 pixels took only around 12 min. For model robustness, three axial view models and three coronal view models were trained and aggregated at a voxel-wise level resulting in a combined model. Predicting the segmentation of one scan with 192 slices by such a combined model took around 90 s using an Intel (Santa Clara, CA, USA) Xeon central processing unit (CPU) (E3-1225v3) and only 6 s when using a GPU.

Table 2 Performance comparison between the accuracy of the automated segmentation achieved by the combined model and the intrarater reliability or interrater reliability, respectively. ↓ indicates that a smaller value represents better performance; *bold p*-values are significant ($p \leq 0.05$)

Metric, median (IQR)	Automated segmentation (AS)	Intrarater reliability	Interrater reliability	*p*-value (AS vs. intrarater)	*p*-value (AS vs. interrater)
VS, in %	95.9 (95.4, 97.2)	94.6 (93.2, 98.4)	89.6 (87.2, 94.1)	0.959	**0.047**
HD95, in mm↓	1.12 (1.12, 1.34)	0.93 (0.71, 1.17)	1.96 (1.54, 2.69)	**0.011**	0.203
DSC, in %	80.0 (78.4, 81.2)	81.8 (80.4, 82.6)	70.5 (69.3, 71.8)	**<0.005**	**<0.005**

VS volumetric similarity, *HD95* 95th percentile of Hausdorff distance, *DSC* Dice similarity coefficient, *IQR* interquartile range

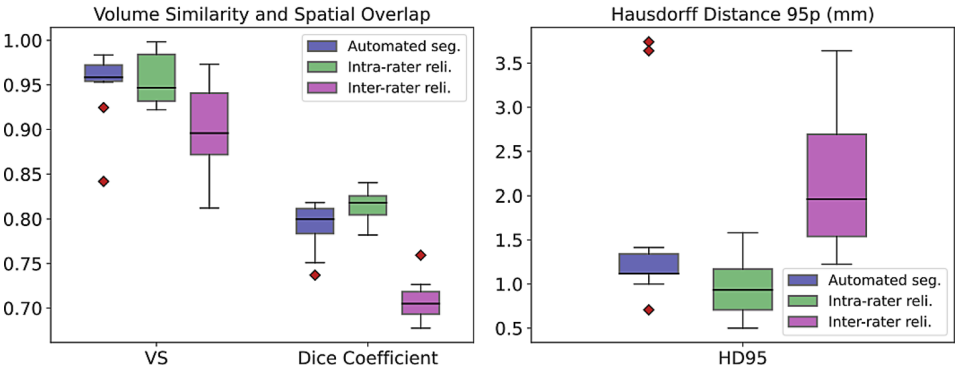


Fig. 4 Segmentation performance of the proposed method on the test set (automated seg.) and comparison to intrarater and interrater reliability (reli.). In comparison with intrarater reliability, automated segmentation is significantly inferior concerning the 95th percentile of the Hausdorff distance (HD95) and Dice coefficient. In comparison with interrater reliability, automated segmentation is significantly superior regarding volumetric similarity (VS) and Dice coefficient (in arbitrary unit, respectively)

Results

Segmentation Accuracy

Three examples of automated claustrum segmentation are shown in Fig. 3.

To assess the accuracy of our combined model for automated claustrum segmentation, we calculated three performance metrics, volumetric similarity (VS), 95th percentile of the Hausdorff distance (HD95), and Dice similarity coefficient (DSC), on the test set and compared its performance with intrarater and interrater reliabilities on the same set (for detailed results see Table 2 and Fig. 4). The proposed method yielded median VS, HD95, and DSC of 95.9%, 1.12 mm, and 80.0%, respectively. Repeated segmentation by the same reader led to median VS, HD95, and DSC of 94.6%, 0.93 mm, and 81.8%, respectively and is referred to as intrarater reliability. Segmentation of the test set by both readers 1 and 2 led to median VS, HD, and DSC of 89.6%, 1.96 mm, and 70.5%, respectively and serves as interrater reliability. Comparing the automated segmentation with intrarater reliability with a Wilcoxon signed-rank test, we found significantly lower HD95 ($p = 0.011$) and higher DSC ($p < 0.005$) for repeated manual segmentation by the same reader. Comparing the automated segmentation with interrater reliability with the same statistical test, the automated segmentation algorithm achieved significantly higher VS ($p = 0.047$) and higher DSC ($p < 0.005$). These results show that the accuracy of our automated segmentation approach is comparable to intrarater reliability with minimally inferior results at HD95 and DSC and that it is superior to interrater reliability in two out of three performance metrics.

Efficiency of Transfer Learning in Comparison with Nontransfer Learning

To evaluate the efficiency of the transfer learning technique (TL), we compared it with the vanilla approach, i.e., training from scratch (non-TL). Internal fivefold cross-validation on the training set was performed with both methods. VS, HD95, DSC and training times were recorded and compared

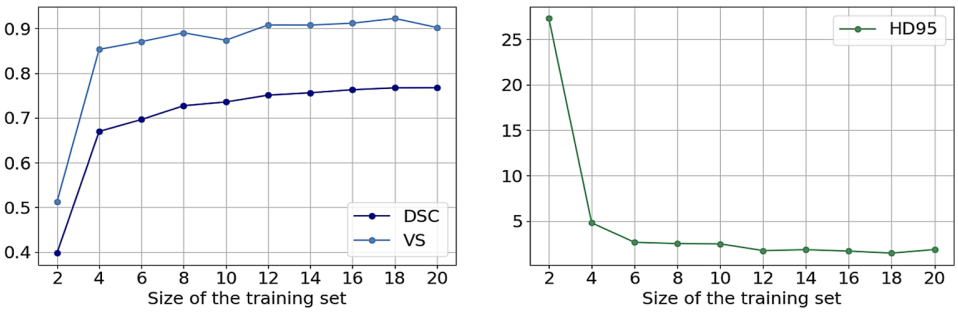


Fig. 5 The left diagram shows volumetric similarity (VS) and Dice similarity coefficient (DSC), both in arbitrary unit, of the test set of models trained with different amounts of training data (measured in scans). The right graph presents the 95th percentile of Hausdorff distance (HD95) in mm of these models. The performance mainly increases till around 12 images in the training set and saturates afterward

(for details, see Table S2 and Fig. S6 in the Online Supplement). The TL method achieved a median VS, HD95, and DSC of 95.3%, 1.06 mm, and 78.9%, respectively, and training took around 90 min. The non-TL approach led to a median VS, HD95, and DSC of 93.5%, 1.00 mm, and 79.4%, respectively, and a training time of 17.5 h. Comparing these results with a Wilcoxon signed-rank test, the TL method showed a significantly superior VS ($p = 0.050$), inferior HD95 ($p = 0.016$), and no significant difference regarding DSC ($p = 0.452$). Concerning the time needed for training, TL was more than 11 times faster than training from scratch. This finding suggests that TL and non-TL achieve comparable performance but TL is far more time efficient.

Data Range Needed for Transfer Learning

To determine how much training data are needed for transfer learning, a model was trained with various training set sizes, i.e., the first model was trained with two scans and the training set was gradually increased with two scans for the following models. The VS, HD95, and DSC were determined on the test set. The model performance improved with increasing training set up to 12 images (Fig. 5). Beyond this size, there only remained a minimal shift of DSC up to 18 images. Surprisingly, even a training set of four scans can reach relatively high scores. This result indicates that transfer learning can deal effectively with a small training set of around 12 scans and their corresponding manual segmentations. Additional results of how much data are needed for non-transfer learning are shown in Fig. S6 in the Online Supplement.

Applicability Assessment on a Large-scale Held-out Correction Set

To test the applicability of the proposed deep-learning-based approach, the model predicted the claustrum in the held-out correction set of 528 scans. Subsequently, we corrected the predictions manually where needed and compared predicted and corrected segmentation by charging VS, HD95, and DSC. The median VS, HD95, and DSC were 98.5%, 0.00 mm, and 97.7% (see Fig. 6), respectively. In total, we found 14 scans of which the DSC of the claustrum segmentation was less than the mean intrarater reliability of 81.8%, corresponding to 2.7% of the whole correction set. In three of these scans, the right claustrum was not detected at all. These subjects, two female and one male neonate, were born in a range of gestational age 26.1–28.7 weeks and scanned between 29.3 and 31 gestational weeks, suggesting an unfavorable impact of very young age on the accuracy of the prediction. A performance comparison between the right and left claustrum is shown in the Online Supplement in Fig. S8 and Table S3.

In a further analysis, we tried to explain the result of the outliers with low performance (DSC < 81.8%). As shown in

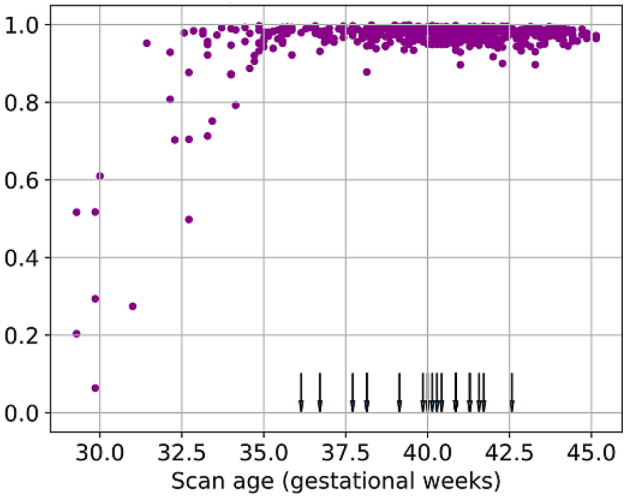


Fig. 7 Dice similarity coefficient (DSC, in arbitrary unit) of 528 manually corrected and initial automated segmentations of right and left claustrum depending on the scan age. The head-down arrows indicate the scan age of the training subjects. Subjects with relatively low segmentation performance are younger than the training samples

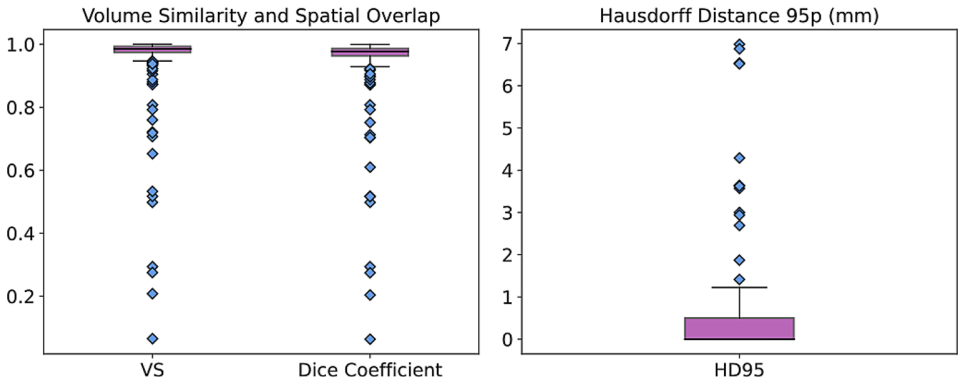


Fig. 6 Volumetric similarity (VS, in arbitrary unit), Dice similarity coefficient (DSC, in arbitrary unit) and 95th percentile of the Hausdorff distance (HD95, in mm) of 528 automated segmentations of the claustrum. Except for several outliers with medium or low accuracy, the majority shows high performance in all three metrics within a small range

Fig. 7, all predictions with low accuracy were obtained in newborns before 35.0 gestational weeks. In subjects older than 35.0 gestational weeks, the combined model reached a high accuracy (DSC >81.8%) in 100% of the scans. Notably, the training subjects were scanned in a range of 36.1–42.6 gestational weeks which presents a domain shift compared to the correction set. Three exemplary young subjects are presented in Fig. S9 in the Online Supplement. This indicates an age-dependent artificial intelligence performance which could be attributed to restricted training samples. Thus, an adjustment of the training samples should improve the performance in young subjects. To test this hypothesis, we replaced two older neonates (scan age around 40 gestational weeks) by two very preterm-born subjects (scan age around 29 gestational weeks) to obtain age stratification in the training set. This led to significantly higher performance in a group of the five young neonates (scan age 29.3–32.7 gestational weeks) with the lowest DSC in Fig. 7 (see Fig. S10) and, surprisingly, also in the original test set (scan age 38.7–42.3 gestational weeks) (see Table S4). To sum up, a scan age stratification of the training set globally improved the model in this developing cohort.

Discussion

This study demonstrated that fully automated claustrum segmentation in T2-weighted neonatal brain MRI is feasible by using deep learning. While the gray matter structure is too small for atlas-based labeling and too intensive for large-scale manual labeling, we successfully implemented a transfer learning (TL) approach building on a previous method for claustrum segmentation in adult brain MRI, leading to segmentation accuracy comparable to intrarater reliability and superior to interrater reliability. The released models and codes will facilitate MRI-based research of the newborn claustrum through automated segmentation. In addition, the presented approach can function as a template for automated segmentation of other intricate structures in the developing neonatal brain or transfer learning to different datasets by published model training and testing code.

The proposed transfer-learning-based method offers high segmentation accuracy. A transfer learning approach fits to our segmentation problem in neonates because DL-based segmentation approaches are more common in adults but not in neonates e.g., amygdala nuclei or hypothalamus [37, 38]. In principle, evidence for the possibility to transfer adult segmentation of specific subcortical regions to neonates was demonstrated. The performance of our segmentation approach was evaluated with three metrics, volumetric similarity (VS), 95th percentile of the Hausdorff distance (HD95) and the Dice similarity coefficient (DSC), on a test set and compared with intrarater and interrater reliability of the same test set. Automated segmentation was partly inferior to intrarater reliability but significantly superior to interrater reliability concerning two scores. In comparison with the prior study of automated adult claustrum segmentation [26], all scores of the neonate claustrum were improved. A possible explanation for this might be the enhanced resolution of newborn MRI/adult MRI of 0.5/1.0 mm isotropic voxel size suggesting that a higher image resolution and a larger volume in voxels lead to higher accuracy. The overall performance level is lower than in comprehensive white or gray matter segmentation reaching a Dice score of about 95% [39]; however, the accuracy accords with observations in other ambiguous and small structures like the hypothalamus and its subnuclei with a Dice score of 51–84% [37]. Altogether, the deep learning method deals with the delicate and variable neonatal claustrum despite a short training set of 20 scans segmented by one rater and outperforms the variability of several human raters, which is especially relevant in large datasets.

When matching TL with non-TL, both options had comparable performance but TL was more time efficient. The methods were optimized individually regarding the number of epochs for training. A second analysis (shown in the Online Supplement) compared the methods with different sizes of the training set with a similar result for larger training sets. With these approaches, a general superiority of TL in terms of our metrics was not certifiable which is consistent with other image segmentation tasks [33]. In the training process, the loss was lower with TL than with non-TL (see Fig. S4 in the Online Supplement) which could be explained by the fact that the Dice loss is not simply confined to the DSC but also represents the certainty of the prediction. To conclude, TL is more time efficient and energy saving than non-TL with stable performance.

We further found that 12 scans for training can be enough to achieve a high model performance. A larger training set hardly improved the accuracy determined with VS, HD95, and DSC. Compared to our previous study, the needed data are much smaller in this neonate project than for adult claustrum segmentation, even after correcting for different image resolutions [26]. Surprisingly, overfitting did not prevent the learning process with small training sets. This could be due to the variability of the images as they come from different layers of the brain. The effect of data augmentation was excluded by testing how much data are needed for models trained without DA. This approach requires more training data for the same performance. We did not test non-DA-non-TL models which would be the exact correlate to the previous adult study. In a large cohort like the dHCP, automated segmentation by deep learning can reduce manual segmentation for the most part as the training and test set are only a small fraction of the whole dataset.

On the question of model applicability, the combined model, an ensemble of three axial and three coronal networks, detected the claustrum correctly in 97.4% of a large held-out correction set. The automated segmentation was compared with manually corrected versions of these predictions and evaluated with VS, HD95, and DSC. The mostly uniform Hausdorff distance of 0.0 mm or 0.5 mm could be attributed to the 95th percentile of this score in conjunction with barely significant adaptations of the predictions. All inadequate predictions with DSC lower than median intrarater reliability were obtained in newborns younger than 35.0 gestational weeks. This result could be explained by the training set which exclusively covered older neonates. Extremely immature neuroanatomy, such as less gyrification or different contrast appearance in MRI than in older neonates, might have distracted our model and resulted in undersegmentation (i.e., false negatives). An age-stratified training set improved the performance in these young subjects and in older neonates. Overall, annotation correction is far more time efficient than manual segmentation from scratch. An automatic selection of subjects that should pass visual control, e.g., due to young age or insufficient detected claustrum volume, could speed up this process further as segmentation in older subjects worked without big mistakes. Consequently, manual correction might be expendable in the latter group. The proposed TL method successfully segments the claustrum with little need for control and correction and enables claustrum analyses in large neonatal cohorts. This facilitates the investigation of the claustrum development and its relation to premature birth. Further investigations are needed to examine the association with other neurodevelopmental disorders, such as schizophrenia and autism spectrum disorders [7].

Despite efficient and accurate automated segmentation, our study has some limitations. First, it is a challenge to precisely define the boundaries of the small and intricate claustrum. Although the dHCP provides a very high isotropic resolution of 0.5 mm and a segmentation protocol structured the process (Online Supplement), the manual segmentation is not perfect because the boundary of specific regions is often ambiguous and its segmentation partly remains subjective, i.e., depends on the rater [37, 40]. This kind of data uncertainty commonly exists in medical image segmentation tasks. One potential solution is to quantify the segmentation uncertainty (e.g., interrater reliability) when building the segmentation model and take the uncertainty of the outcome into account for the downstream analysis (Sect. Segmentation Accuracy). Second, all training images were segmented by one rater. This improves the uniformity of segmentations but could also lead to a bias of the model. Further analyses with two or more raters would be necessary to appraise this impact. Third, the model training was limited to a small dataset that did not cover the whole age range of the dHCP or all neonatal stages of development, which presumably dropped the accuracy, especially in early premature newborns. The model still facilitates manual work in the affected subjects but a strong visual control is important.

In conclusion, this study presented a deep learning approach for automated claustrum segmentation in human neonatal brain MRI. We evaluated the accuracy, compared transfer and non-transfer learning, analyzed how much data are needed for transfer learning and assessed the applicability of the proposed method including a model enhancement by age-stratified training. We conclude that 1) transfer learning is a bit inferior to intrarater reliability but superior to interrater reliability, 2) transfer learning shows similar performance to non-transfer learning and is more time efficient, 3) the prediction accuracy stabilizes with a training set above 12 scans and 4) the combined model applies to a large cohort with predominantly accurate results. The implementation codes are available on *GitHub* to the research community.

Supplementary Information The online version of this article (https://doi.org/10.1007/s00062-021-01137-8) contains supplementary material, which is available to authorized users.

Acknowledgements Data were provided by the developing Human Connectome Project, KCL-Imperial-Oxford Consortium funded by the European Research Council under the European Union Seventh Framework Programme (FP/2007-2013)/ERC Grant Agreement no. [319456]. We are grateful to the families who generously supported this trial.

Funding This study is supported by the Deutsche Forschungsgemeinschaft (SO 1336/1-1 to Christian Sorg), German Federal Ministry of Education and Science (BMBF 01ER0803 to Christian Sorg) and the Kommission für Klinische Forschung, Technische Universität München (KKF 8765162 to Christian Sorg). Hongwei Bran Li is supported by Forschungskredit (Grant NO. FK-21-125) from University of Zurich. Open Access funding was enabled and organized by Projekt DEAL.

Conflict of interest A. Neubauer, H.B. Li, J. Wendt, B. Schmitz-Koep, A. Menegaux, D. Schinz, B. Menze, C. Zimmer, C. Sorg and D.M. Hedderich declare that they have no competing interests.

References

1. Puelles L. Development and evolution of the claustrum. In: Smythies JR, Edelstein L, Ramachandran VS, editors. The claustrum: structural, functional, and clinical neuroscience. Amsterdam: Elsevier Academic Press; 2014. pp. 119–76.
2. Kowiański P, Dziewiatkowski J, Kowiańska J, Moryś J. Comparative anatomy of the claustrum in selected species: A morphometric analysis. Brain Behav Evol. 1999;53:44-54.
3. Crick FC, Koch C. What is the function of the claustrum? Philos Trans R Soc Lond B Biol Sci. 2005;360:1271-9.
4. Brown SP, Mathur BN, Olsen SR, Luppi PH, Bickford ME, Citri A. New Breakthroughs in Understanding the Role of Functional Interactions between the Neocortex and the Claustrum. J Neurosci. 2017;37:10877-81.
5. Goll Y, Atlan G, Citri A. Attention: the claustrum. Trends Neurosci. 2015;38:486-95.
6. Mathur BN. The claustrum in review. Front Syst Neurosci. 2014;8. https://doi.org/10.3389/fnsys.2014.00048/full.
7. Smith JB, Lee AK, Jackson J. The claustrum. Curr Biol. 2020;30: R1401-6.
8. White MG, Panicker M, Mu C, Carter AM, Roberts BM, Dharmasri PA, Mathur BN. Anterior Cingulate Cortex Input to the Claustrum Is Required for Top-Down Action Control. Cell Rep. 2018;22:84-95.
9. Krimmel SR, White MG, Panicker MH, Barrett FS, Mathur BN, Seminowicz DA. Resting state functional connectivity and cognitive task-related activation of the human claustrum. Neuroimage. 2019;196:59-67.
10. Arrigo A, Mormina E, Calamuneri A, Gaeta M, Granata F, Marino S, Anastasi GP, Milardi D, Quartarone A. Inter-hemispheric Claustral Connections in Human Brain: A Constrained Spherical Deconvolution-Based Study. Clin Neuroradiol. 2017;27:275-81.
11. Milardi D, Bramanti P, Milazzo C, Finocchio G, Arrigo A, Santoro G, Trimarchi F, Quartarone A, Anastasi G, Gaeta M. Cortical and subcortical connections of the human claustrum revealed in vivo by constrained spherical deconvolution tractography. Cereb Cortex. 2015;25:406-14.
12. Binks D, Watson C, Puelles L. A Re-evaluation of the Anatomy of the Claustrum in Rodents and Primates-Analyzing the Effect of Pallial Expansion. Front Neuroanat. 2019;13:34.
13. Watson C, Puelles L. Developmental gene expression in the mouse clarifies the organization of the claustrum and related endopiriform nuclei. J Comp Neurol. 2017;525:1499-508.
14. Brittain PJ, Froudist Walsh S, Nam KW, Giampietro V, Karolis V, Murray RM, Bhattacharyya S, Kalpakidou A, Nosarti C. Neural compensation in adulthood following very preterm birth demonstrated during a visual paired associates learning task. Neuroimage Clin. 2014;6:54-63.
15. Bruguier H, Suarez R, Manger P, Hoerder-Suabedissen A, Shelton AM, Oliver DK, Packer AM, Ferran JL, García-Moreno F, Puelles L, Molnár Z. In search of common developmental and evolutionary origin of the claustrum and subplate. J Comp Neurol. 2020;528:2956-77.
16. Kanold PO, Luhmann HJ. The subplate and early cortical circuits. Annu Rev Neurosci. 2010;33:23-48.
17. Kinney HC, Haynes RL, Xu G, Andiman SE, Folkerth RD, Sleeper LA, Volpe JJ. Neuron deficit in the white matter and subplate in periventricular leukomalacia. Ann Neurol. 2012;71:397-406.
18. McClendon E, Shaver DC, Degener-O'Brien K, Gong X, Nguyen T, Hoerder-Suabedissen A, Molnár Z, Mohr C, Richardson BD, Rossi DJ, Back SA. Transient Hypoxemia Chronically Disrupts Maturation of Preterm Fetal Ovine Subplate Neuron Arborization and Activity. J Neurosci. 2017;37:11912-29.
19. McQuillen PS, Ferriero DM. Perinatal subplate neuron injury: implications for cortical development and plasticity. Brain Pathol. 2005;15:250-60.
20. Volpe JJ. Dysmaturation of premature brain: importance, cellular mechanisms, and potential interventions. Pediatr Neurol. 2019;95:42–66.
21. Volpe JJ. Subplate neurons--missing link in brain injury of the premature infant? Pediatrics. 1996;97:112-3.
22. Hedderich DM, Menegaux A, Li H, Schmitz-Koep B, Stämpfli P, Bäuml JG, Berndt MT, Bäuerlein FJB, Grothe MJ, Dyrba M, Avram M, Boecker H, Daamen M, Zimmer C, Bartmann P, Wolke D, Sorg C. Aberrant Claustrum Microstructure in Humans after Premature Birth. Cereb Cortex. 2021;31:5549-59.
23. Hoerder-Suabedissen A, Oeschger FM, Krishnan ML, Belgard TG, Wang WZ, Lee S, Webber C, Petretto E, Edwards AD, Molnár Z. Expression profiling of mouse subplate reveals a dynamic gene network and disease association with autism and schizophrenia. Proc Natl Acad Sci USA. 2013;110:3555-60.
24. Berman S, Schurr R, Atlan G, Citri A, Mezer AA. Automatic Segmentation of the Dorsal Claustrum in Humans Using in vivo High-Resolution MRI. Cereb Cortex Commun. 2020;1:tgaa062.
25. Albishri AA, Shah SJH, Schmiedler A, Kang SS, Lee Y. Automated human claustrum segmentation using deep learning technologies. 2019. http://arxiv.org/abs/1911.07515. Accessed 21 May 2021.
26. Li H, Menegaux A, Schmitz-Koep B, Neubauer A, Bäuerlein FJB, Shit S, Sorg C, Menze B, Hedderich D. Automated claustrum segmentation in human brain MRI using deep learning. Hum Brain Mapp. 2021;42:5862-72.
27. Hughes EJ, Winchman T, Padormo F, Teixeira R, Wurie J, Sharma M, Fox M, Hutter J, Cordero-Grande L, Price AN, Allsop J, Bueno-Conde J, Tusor N, Arichi T, Edwards AD, Rutherford MA, Counsell SJ, Hajnal JV. A dedicated neonatal brain imaging system. Magn Reson Med. 2017;78:794-804.
28. Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng. 2010;22:1345–59.
29. Makropoulos A, Robinson EC, Schuh A, Wright R, Fitzgibbon S, Bozek J, Counsell SJ, Steinweg J, Vecchiato K, Passerat-Palmbach J, Lenz G, Mortari F, Tenev T, Duff EP, Bastiani M, Cordero-Grande L, Hughes E, Tusor N, Tournier JD, Hutter J, Price AN, Teixeira RPAG, Murgasova M, Victor S, Kelly C, Rutherford MA, Smith SM, Edwards AD, Hajnal JV, Jenkinson M, Rueckert D. The developing human connectome project: A minimal processing pipeline for neonatal cortical surface reconstruction. Neuroimage. 2018;173:88-112.
30. Carney O, Hughes E, Tusor N, Dimitrova R, Arulkumaran S, Baruteau KP, Collado AE, Cordero-Grande L, Chew A, Falconer S, Allsop JM, Rueckert D, Hajnal J, Edwards AD, Rutherford M. Incidental findings on brain MR imaging of asymptomatic term neonates in the Developing Human Connectome Project. EClinicalMedicine. 2021;38:100984.
31. Cordero-Grande L, Teixeira RPAG, Hughes EJ, Hutter J, Price AN, Hajnal JV. Sensitivity encoding for aligned multishot magnetic resonance reconstruction. IEEE Trans Comput Imaging. 2016;2: 266–80.
32. Yushkevich PA, Piven J, Hazlett HC, Smith RG, Ho S, Gee JC, Gerig G. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. Neuroimage. 2006;31:1116-28.
33. Karimi D, Warfield SK, Gholipour A. Critical assessment of transfer learning for medical image segmentation with fully convolutional neural networks. 2020. http://arxiv.org/abs/2006.00356. Accessed 19 Aug 2021.
34. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, editors. Medical image computing and

computer-assisted intervention—MICCAI 2015. Cham: Springer; 2015. pp. 234–41.
35. Milletari F, Navab N, Ahmadi SA. V-net: fully convolutional neural networks for volumetric medical image segmentation. 2016 fourth international conference on 3D vision (3DV). 2016. pp. 565–71.
36. Ma J, Chen J, Ng M, Huang R, Li Y, Li C, Yang X, Martel AL. Loss odyssey in medical image segmentation. Med Image Anal. 2021;71:102035.
37. Billot B, Bocchetta M, Todd E, Dalca AV, Rohrer JD, Iglesias JE. Automated segmentation of the hypothalamus and associated subunits in brain MRI. Neuroimage. 2020;223:117287.
38. Saygin ZM, Kliemann D, Iglesias JE, van der Kouwe AJW, Boyd E, Reuter M, Stevens A, Van Leemput K, McKee A, Frosch MP, Fischl B, Augustinack JC; Alzheimer's Disease Neuroimaging Initiative. High-resolution magnetic resonance imaging reveals nuclei of the human amygdala: manual segmentation to automatic atlas. Neuroimage. 2017;155:370-82.
39. Gabr RE, Coronado I, Robinson M, Sujit SJ, Datta S, Sun X, Allen WJ, Lublin FD, Wolinsky JS, Narayana PA. Brain and lesion segmentation in multiple sclerosis using fully convolutional neural networks: A large-scale study. Mult Scler. 2020;26:1217-26.
40. Fischl B, Salat DH, Busa E, Albert M, Dieterich M, Haselgrove C, van der Kouwe A, Killiany R, Kennedy D, Klaveness S, Montillo A, Makris N, Rosen B, Dale AM. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron. 2002;33:341-55.

5 Unified multi-contrast MR neuroimage synthesis and its clinical validation

This chapter has been published as one **peer-reviewed conference publication** and one **peer-reviewed journal publication**:

[1] **H. Li***, J. C. Paetzold*, A. Sekuboyina, F. Kofler, J. Zhang, J. S. Kirschke, B. Wiestler, and B. Menze. "DiamondGAN: unified multi-modal generative adversarial networks for MRI sequences synthesis". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (**MICCAI**)*. Springer. 2019, pp. 795–803

[2] T. Finck*, **H. Li***, L. Grundl, P. Eichinger, M. Bussas, M. Mühlau, B. Menze, and B. Wiestler. "Deep-learning generated synthetic double inversion recovery images improve multiple sclerosis lesion detection". In: ***Investigative Radiology*** 55.5 (2020), pp. 318–323

Synopsis: The above two works develop a unified synthesis framework for multi-contrast MR image synthesis based on generative adversarial networks and validate it in a clinical setting. Specifically, publication ♯1 develops a conditional synthesis approach based on generative adverarial network and allows for flexible training and arbitrary missing contrast synthesis. Publication ♯2 implements the developed technique, generates synthetic double inversion recovery (synthDIR) images and compare their diagnostic performance to conventional sequences in patients with multiple sclerosis (MS). We observe that the generated DIR images improve lesion depiction compared with the use of standard modalities.

Contributions of thesis author: algorithm design and implementation, computational experiments and composition of manuscript.

DiamondGAN: Unified Multi-modal Generative Adversarial Networks for MRI Sequences Synthesis

Hongwei Li[1], Johannes C. Paetzold[1], Anjany Sekuboyina[1,2], Florian Kofler[1], Jianguo Zhang[3,4(✉)], Jan S. Kirschke[2], Benedikt Wiestler[2], and Bjoern Menze[1,5]

[1] Department of Informatics, Technical University of Munich, Munich, Germany
{hongwei.li,bjoern.menze}@tum.de
[2] Department of Neuroradiology, Klinikum rechts der Isar, Munich, Germany
[3] Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, China
jgzhang@ieee.org
[4] Shenzhen Institute of Artificial Intelligence and Robotics for Society, Shenzhen, China
[5] Institute for Advanced Study, Technical University of Munich, Munich, Germany

Abstract. Synthesizing MR imaging sequences is highly relevant in clinical practice, as single sequences are often missing or are of poor quality (e.g. due to motion). Naturally, the idea arises that a target modality would benefit from multi-modal input, as proprietary information of individual modalities can be synergistic. However, existing methods fail to scale up to multiple non-aligned imaging modalities, facing common drawbacks of complex imaging sequences. We propose a novel, scalable and multi-modal approach called *DiamondGAN*. Our model is capable of performing flexible non-aligned cross-modality synthesis and data infill, when given multiple modalities or any of their arbitrary subsets, learning structured information in an end-to-end fashion. We synthesize two MRI sequences with clinical relevance (i.e., double inversion recovery (DIR) and contrast-enhanced T1 (T1-c)), reconstructed from three common sequences. In addition, we perform a multi-rater visual evaluation experiment and find that trained radiologists are unable to distinguish synthetic DIR images from real ones.

1 Introduction

In clinical practice, magnetic resonance imaging (MRI) datasets often consists of high-dimensional image volumes with multiple imaging protocols and repeated scans acquired at multiple time points. Given the multiplicity of possible sequence parameters, protocols largely vary depends on the imaging centers,

H. Li and J. C. Paetzold—Equal contribution.

D. Shen et al. (Eds.): MICCAI 2019, LNCS 11767, pp. 795–803, 2019.
https://doi.org/10.1007/978-3-030-32251-9_87

hindering their comparability. This often leads to repeated exams or severely limits the clinical information that can be drawn from those MRI studies. Particularly, in the case of multiple sclerosis, longitudinal comparisons of MRI studies are the main reason for treatment decisions and existing lesion quantification tools require complete identical modalities at multiple time points. Potentially, cross-modality image synthesis technique can resolve those obstacles through efficient data infilling and re-synthesis.

Recently, generative adversarial networks (GANs) have been applied in translating MRI sequences, positron emission tomography (PET) and computed tomography (CT) images. Most of them are one-to-one cross-modality synthesis approaches, for example, PET [12] synthesis and MRI sequences translation [3]. A recent multi-modal synthesis method [10] has limited scalability because the input and output modalities are required to be spatially aligned. Although there are several multi-domain translation algorithms [2] in the computer vision community, these approaches design one-to-multiple domain translation but do not model the multiple-to-one domain mapping. Especially in medical images synthesis, *multiple-to-one* cross-modality mapping is highly relevant as proprietary information of individual and non-aligned modalities can be synergistic.

There are three main challenges in the scenario of multi-modal cross-modality medical image synthesis: (1) the input and target modalities are assumed to be *not* spatially-aligned because registration methods for aligning multiple modalities may fail, restricting the applicability of conventional regression approaches. (2) input modalities may be missing due to different clinical settings between centers, thus a traditional regression-based data infill would be restricted to the smallest uniform subset or rely on iterative data infill methods. (3) existing approaches have limited scalability, e.g. in a *Cycle-GAN* [14] setting, one would therefore have to train individual models for possible combinations of the input modalities.

Contributions (1) We propose *DiamondGAN*, which is a unified, scalable multi-modal generative adversarial network. It learns the multiple-to-one cross-modality mapping among non-aligned modalities using only a pair of generators and discriminators, optimized with a multi-modal cycle-consistency loss function. (2) We provide both qualitative and quantitative results on two clinically-relevant MRI sequences synthesis tasks, showing *DiamondGAN's* superiority over baseline models. (3) We present the results of extensive visual evaluation, performed by fourteen experienced radiologists to confirm the quality of synthetic images.

2 Methodology

2.1 Multi-modal Cross-Modality Synthesis

Given an input set of n modalities: $X = \{x_i | i = 1, ..., n\}$ and a target modality T. Our goal is to learn a generator G that learns mappings from multiple input modalities to one target modality. We assume that (1) all the modalities,

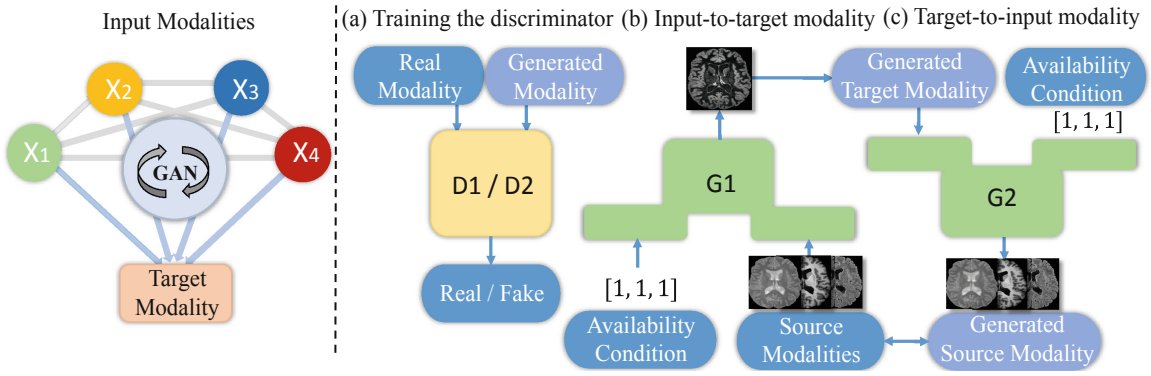


Fig. 1. Left: The high-level idea behind *DiamondGAN*, which is capable of learning mappings between any subset of multiple input modalities (X) to a target modality in a single model. This mapping represents a diamond-shape topology. Right: Overview of *DiamondGAN*. It consists of two modules, a pair of discriminators D and a pair of generators G. (a) *D1* and *D2* learn to distinguish between the real and synthetic images from multi-modal input and the target output respectively. (b) *G1* takes both multi-modal input and the condition as input and generates a target modality. The condition c is a binary vector: $c = \{c_1, c_2, ..., c_n\}$, where c_i indicates the corresponding input modality as available (1) or not (0). It is spatially replicated and concatenated with the input modalities in the feature level. (c) *G2* tries to generate the original modalities from the synthetic target modality given the original availability condition.

i.e., X and T, are *not* spatially-aligned because it is rather difficult to obtain *strictly* spatially-aligned images as mentioned in Sect. 1; (2) the input modalities can be any subset of X, denoted as X' during the training and inference stages as some modalities of a subject may be missing in clinical practice.

We enforce G to be capable of translating any subset X' into a target modality T using a condition c which indicates the presence of the input modalities, i.e., $G(X', c) \rightarrow T$. This condition handles the missing modality issue and makes it a scalable model in both the training and the inference stages. We further introduce a multi-modal cycle-consistency loss to handle the "non-aligned modalities" issue among the input and output. Figure 1 illustrates the main idea of our proposed approach. We regularly generate the condition c and the corresponding multi-modal data X_c of all possible combinations, so that G learns to flexibly translate the arbitrary multi-modal input. As mentioned in the caption of Fig. 1, we use an *availability* condition to serve as an indicator of the input modalities. It is spatially replicated to the image size ($1 \times H \times W$) and is a part of the two-stream network input. In the case of 3 modalities as the input, the condition $c = [1, 1, 1]$ would indicate that every input modality is given.

Multi-modal Reconstruction Loss. We aim to train G to guarantee that a generated target modality preserves the content of its input modalities. The input modalities are assumed to be not spatially aligned or not from the same subject as mentioned above. In this situation, the traditional cycle loss [14] as well as the regression loss [5] would fail to tackle the multi-modal and

non-alignment issues. To alleviate the two problems, we extend the traditional cycle-consistency loss [14] to a multi-modal one. Specifically, we concatenate the source modalities into a multi-channel input and define a multi-channel output as the target modality. We then simultaneously train two generators $G_1 : X \rightarrow T$ and $G_2 : T \rightarrow X$ in a cycle-consistency fashion. Please note that the output target modality is in multiple channels which correspond to the input modalities. The loss function of the generator is defined as:

$$\mathcal{L}_{rec} = \mathbb{E}_{X,T,c}[||X - G_2(G_1(X,c),c)||_1 + ||T - G_1(G_2(T,c),c)||_1] \quad (1)$$

Adversarial Loss. To make the generated images indistinguishable from real images, we adopt an adversarial loss:

$$\begin{aligned} \mathcal{L}_{adv} = {} & \mathbb{E}_{X,T}\{log\ [D_1(X) \cdot D_2(T)]\} \\ & + \mathbb{E}_{X,T,c}\{log\ [(1 - D_2(G_1(X,c))) \cdot (1 - D_1(G_2(T,c)))]\} \end{aligned} \quad (2)$$

where G_1 generates a target modality $G_1(X, c)$ conditioned on the presence of input modalities X, while D_1 tries to distinguish between real input modalities and generated ones. Similarly, G_2 generates the original input modalities $G_2(T, c)$ conditioned on the presence of original input modalities X and D_2 tries to distinguish between the real target modality and the generated one. The generators try to minimize this objective, while the discriminators to maximize it.

Full Objective. The objective functions to optimize D and G respectively are

$$\mathcal{L}_D = -\mathcal{L}_{adv}; \quad \mathcal{L}_G = \mathcal{L}_{adv} + \lambda_{rec}\mathcal{L}_{rec} \quad (3)$$

where λ_{rec} is the hyper-parameter that balances the reconstruction loss and adversarial loss.

2.2 Implementation

Two-Stream Network Architecture. To leverage the information from both input modalities and corresponding availability conditions, we build a two-stream network architecture based on the popular encoder-decoder network [6]. It takes the multi-modal images and condition as two inputs and merges them in the feature level. This network contains stride-2 convolutions, residual blocks [4] and fractionally-strided convolutions (1/2 stride). We use 6 blocks for the input size of $N \times H \times W$, where N, H and W are the number of modalities, height and width of the images respectively. The input and availability conditions pass through two encoders and are merged in the last feature layer before the decoder. *PatchGANs* [6] is used for the discriminator network, which classifies the patch feature maps to real or fake, instead of using a fully-connected layer.

Training Details. We apply two recent techniques to stabilize the training of the model. First, for $\mathcal{L}_{adv}$ (Eq. 2), we replace the negative log likelihood objective by a least-squares loss [9]. Second, to reduce the model oscillation, we update the discriminators using a history of generated images rather than the ones produced by the latest generators, as proposed in [11]. Thus we put the 25 previously generated images in an image buffer. We set $\lambda_{rec} = 10$ in Eq. 3 for all the experiments. We use the Adam solver [7] with a batch size of 5. All networks were trained from scratch with a learning rate of 0.0002 and for 20 epochs. When given n input modalities, for each epoch the parameters in both generator and discriminator are updated for $2^n - 1$ times given $2^n - 1$ training subsets of input modalities excluding empty set. The implementations of our model are available in https://github.com/hongweilibran/DiamondGAN.

2.3 Visual Rating and Evaluation Protocol

Quantitative evaluation of generated images in terms of standard scores for errors and correlation remains a debatable task [1]. Additionally, the evaluation with common metrics such as PSNR and MAE [13] would not tell us to whether the algorithm captures clinically relevant small substructures. Therefore, we strive to get experts' estimates of the image quality. We design a multi-rater quality evaluation experiment. Neuro-radiologists rated the images in a browser-application. In each trial, they were provided with two images. On the left side, one real source image of a T1 or Flair images is presented. On the other side, a paired image of the target modality is shown which is either a real image or a generated one. The displayed paired images were randomly chosen in the pool of generated images and real ones. This particular setup enables the experts to identify very small inconsistency or implausibility between the two images immediately. For evaluation, the experts were asked to rate the plausibility of the image on the right based on the real image on the left, to assign a 6-star rating, where 6 stars denoted a perfectly plausible image and 1 star a completely implausible image. The images were presented in 280 trials.

3 Experiments

Datasets. *Dataset 1* consists of 65 scans of patients with MS lesions from a local hospital, acquired with a multi-parametric protocol, which includes co-registered Flair, T1, T2, double inversion recovery (DIR) and contrast-enhanced T1 (T1-c) after skull-stripping. The first three modalities are common modalities in most MS lesion exams. DIR is a MRI pulse sequence, which suppresses signal from the cerebrospinal fluid and the white matter, enhancing the inflammatory lesion. T1-c is a MRI sequence which requires a paramagnetic contrast agent (usually gadolinium) that reduces the T1 relaxation time and thereby increases the signal intensity. Synthesizing DIR and T1-c is of clinical relevance because it can substantially reduce medical costs. We mainly report our result on *Dataset 1*. Additional *Dataset 2* is used to demonstrate that our approach can work on multiple datasets with incomplete and non-aligned modalities. It is a part of the public MICCAI-WMH dataset

[8], and includes 40 subjects with two modalities (Flair and T1). 2D axial slices are used for training the network. All the slices are cropped or padded to a uniform size of 240 × 240 and intensity values are rescaled to [−1, 1].

Reconstructing DIR and T1-c from Common Modalities. We perform two image synthesis tasks on two clinically-relevant MRI sequences (DIR and T1-c), using three common modalities (i.e., Flair, T1 and T2). We separate the *Dataset 1* into a training set, a validation set and a test set, resulting in 30 scans (2015 slices for each modality) for training and 35 scans for testing (2100 slices for each modality). To obtain the optimal hyper-parameters of the model, we use 5 out of the 30 training scans as a validation set. A common approach for quantitative evaluation of medical GAN images is to calculate relative errors and signal to noise ratio between the synthetic image and the real image [13]. Table 1 shows the results of peak signal-to-noise ratio (PSNR) and mean absolute error (MAE) by comparing the synthetic images and real T1-c and DIR images. For the synthetic DIR and T1-c images, we report the highest PSNR and the lowest MAE for a combined T1+T2+Flair input to our model. In the DIR synthesis experiment, the listed scores of using multiple inputs to our GAN are comparable (MAE 0.058-0.065). Whereas, the scores for single inputs are substantially worse (MAE 0.073-0.084). For the T1-c synthesis task, we find that any combination of multi-modal inputs involving the T1 modality (MAE 0.045-0.048) results in better scores compared to other inputs. This indicates that our model successfully extracts the relevant information, as T1-c is a T1 scan with a contrast enhancing agent. For comparison, we implement *CycleGAN* [14] to perform one-to-one cross-modality synthesis, the best results of *CycleGAN* are listed in Table 1. For DIR synthesis, using Flair images as the input of *CycleGAN* achieves the highest PSNR and lowest MAE while for T1-c, using T1 as the input gets the best performance. The proposed model outperforms *CycleGAN* in both

Table 1. Quantitative evaluation of our generated images compared to the real DIR and T1-c image using PSNR and MAE as evaluation metrics. Results show that the generated images benefit from a multi-modal input. ↑ indicates that higher values corresponds to better image qualities.

	DIR $_{PSNR}$↑	DIR $_{MAE}$↓	T1-c $_{PSNR}$↑	T1-c $_{MAE}$↓
CycleGAN [14]	17.34	0.068	20.36	0.045
$DiamonGAN_{T1}$	15.46	0.084	20.21	0.048
$DiamonGAN_{T2}$	15.99	0.073	19.34	0.054
$DiamonGAN_{Flair}$	16.16	0.078	17.15	0.068
$DiamonGAN_{T1+T2}$	17.41	0.065	20.75	0.046
$DiamonGAN_{T2+Flair}$	18.58	0.059	19.78	0.051
$DiamonGAN_{T1+Flair}$	18.02	0.062	20.40	0.047
$DiamonGAN_{T1+T2+Flair}$	**18.63**	**0.058**	**20.86**	**0.045**

tasks. We further replace a part of the training Flair and T1 images in *Dataset 1* with images from *Dataset 2* (totally 794 images for each modality) and we find the result on same testing set is comparable to using the original *Dataset 1*.

Wilconxon signed-rank tests are conducted on the PSNR and MAE pairs generated by *DiamondGAN* (with 3 modalities) and *CycleGAN* respectively. Although the improvements of PSNR and MAE look small in whole image level, they are statistically significant (p-value < 0.0001) in the case of DIR in Table 1. This improvement is highly relevant for biomaker synthesis and for pathological evaluation especially in the case of MS lesions with small volumes. The samples of synthetic T1-c and DIR images are shown in Fig. 2.

Visual Evaluations by Neuroradiologists. Fourteen neuro-radiologists with median 5+ years of professional experience participated. Each evaluated 210 synthetic images and 70 original images. The 210 synthetic images are generated enforcing 6 different input conditions in which each condition includes 35 samples. The rating results of the 14 raters are averaged and the box plots of the results are shown in Fig. 3. For the synthesis of T1-c images, we found that three multi-modal combinations (i.e., *T1*, *T1+Flair* and *T1+T2+Flair*) gave comparable results, while the ones based solely on a Flair were consistently rated as implausible. The plausibility of DIR images synthesized with *T1+T2+Flair* input was rated on average 0.83 stars higher than that with solely T1 input. This is plausible as the DIR is a complex sequence containing proprietary information, its synthesis thus benefits from multiple input sources. For the synthetic

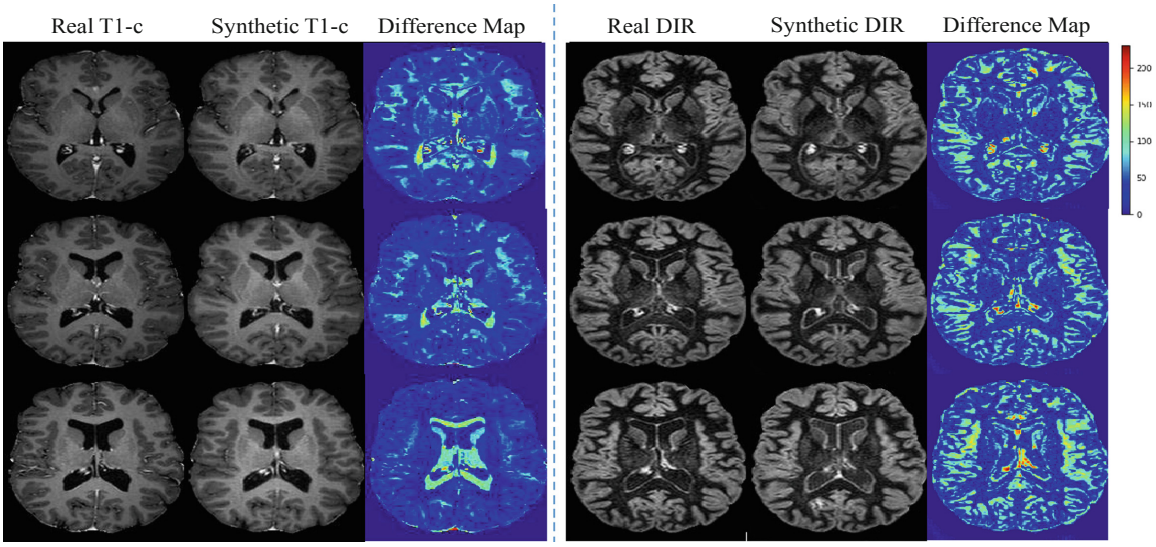


Fig. 2. Samples of synthetic T1-c and DIR images given the combination of T1, T2 and Flair modalities. Difference images are generated and visualized in heat maps. The synthetic images preserve the tissue contrast and the anatomy information. However, we find more differences in synthetic DIR images than in synthetic T1-c ones, especially around the brain boundary. This could be due to the alignment error by registration methods.

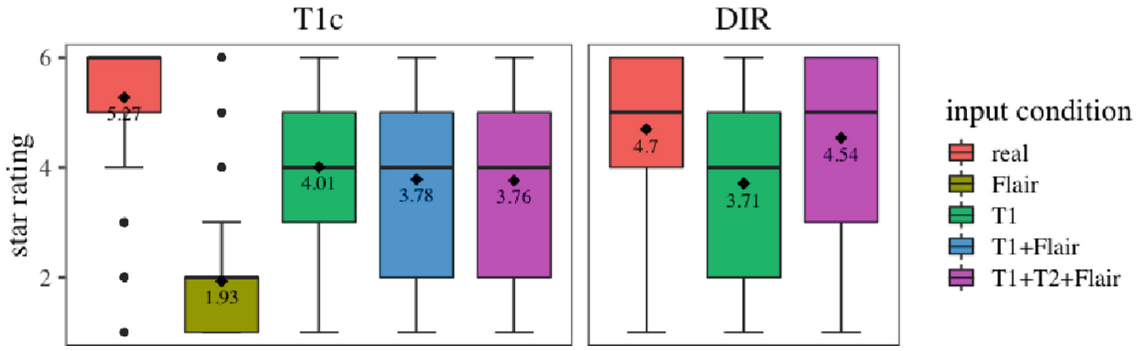


Fig. 3. Box plots showing the rating scores of synthetic images and real ones for T1-c modality on the left and DIR modality on the right. The means are shown as black numbers. *DiamondGAN* achieves comparable plausibility levels for the DIR modality.

images with *T1+T2+Flair* input, the experts assigned an identical rating to the synthetic and original images (4.54 stars *vs* 4.7 stars).

We conduct Wilcoxon rank-sum tests on the paired rating scores of synthetic and real images from 14 raters on 6 conditions which results in 6 pairs of 14 observations. Results show that the pair of rating scores on synthetic DIR images by *T1+T2+Flair* input and real DIR images are not significantly different (p-value $= 0.1432$) while all other pairs are significantly different (p-values < 0.0001). This demonstrates that trained radiologists are unable to distinguish our synthetic DIR images from real ones. Furthermore, the experts ratings for the individual conditions of synthetic images are in agreement with the metrical evaluation in Table 1. For T1-c synthesis, the PSNR and MAE scores are consistently good when T1 modality is fed to *DiamondGAN*.

4 Conclusion and Discussion

This work introduces a novel approach for multi-modal medical image synthesis, with extensive multi-rater experiments and statistical tests. This multi-modal approach allows us to mine the structured information inside the existing extensive MRI sequences. Pathological evaluation is the ultimate goal of this work. Our approach is evaluated by clinical partners who contributed the datasets. We compared synthetic DIR sequence with conventional FLAIR sequence in a MS lesions detection task in a cohort study. The proposed *DiamondGAN* has the potential to reduce medical costs in clinical practice.

Acknowledgement. This work is support by Technische Universität München - Institute for Advanced Study, funded by the German Excellence Initiative and European Union 7^{th} Framework Programme under grant agreement No. 291763. HL and BW are supported by the funding from Zentrum Digitalisierung Bayern.

References

1. Borji, A.: Pros and cons of gan evaluation measures. Comput. Vis. Image Underst. **179**, 41–65 (2019)
2. Choi, Y., et al.: StarGAN: unified generative adversarial networks for multi-domain image-to-image translation. In: CVPR, pp. 8789–8797 (2018)
3. Dar, S.U., et al.: Image synthesis in multi-contrast MRI with conditional generative adversarial networks. IEEE Trans. Med. Imaging (2019)
4. He, K., et al.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)
5. Isola, P., et al.: Image-to-image translation with conditional adversarial networks. In: CVPR, pp. 1125–1134 (2017)
6. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 694–711. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_43
7. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
8. Kuijf, H.J., et al.: Standardized assessment of automatic segmentation of white matter hyperintensities; results of the WMH segmentation challenge. IEEE Trans. Med. Imaging (2019)
9. Mao, X., et al.: Least squares generative adversarial networks. In: CVPR, pp. 2794–2802 (2017)
10. Sharma, A., Hamarneh, G.: Missing MRI pulse sequence synthesis using multi-modal generative adversarial network. arXiv preprint arXiv:1904.12200 (2019)
11. Shrivastava, A., et al.: Learning from simulated and unsupervised images through adversarial training. In: CVPR, pp. 2107–2116 (2017)
12. Wang, Y., et al.: 3D conditional generative adversarial networks for high-quality PET image estimation at low dose. Neuroimage **174**, 550–562 (2018)
13. Welander, P., et al.: Generative adversarial networks for image-to-image translation on multi-contrast MR images-a comparison of cyclegan and unit. arXiv preprint arXiv:1806.07777 (2018)
14. Zhu, J.Y., et al.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: CVPR, pp. 2223–2232 (2017)

Deep-Learning Generated Synthetic Double Inversion Recovery Images Improve Multiple Sclerosis Lesion Detection

Tom Finck, MD, Hongwei Li, PhD,† Lioba Grundl, MD,* Paul Eichinger, MD,* Matthias Bussas, PhD,‡§ Mark Mühlau, MD,‡§ Bjoern Menze, PhD,† and Benedikt Wiestler, MD*†*

Objectives: The aim of the study was to implement a deep-learning tool to produce synthetic double inversion recovery (synthDIR) images and compare their diagnostic performance to conventional sequences in patients with multiple sclerosis (MS).

Materials and Methods: For this retrospective analysis, 100 MS patients (65 female, 37 [22–68] years) were randomly selected from a prospective observational cohort between 2014 and 2016. In a subset of 50 patients, an artificial neural network (*DiamondGAN*) was trained to generate a synthetic DIR (synthDIR) from standard acquisitions (T1, T2, and fluid-attenuated inversion recovery [FLAIR]). With the resulting network, synthDIR was generated for the remaining 50 subjects. These images as well as conventionally acquired DIR (trueDIR) and FLAIR images were assessed for MS lesions by 2 independent readers, blinded to the source of the DIR image. Lesion counts in the different modalities were compared using a Wilcoxon signed-rank test, and interrater analysis was performed. Contrast-to-noise ratios were compared for objective image quality.

Results: Utilization of synthDIR allowed to detect significantly more lesions compared with the use of FLAIR images (31.4 ± 20.7 vs 22.8 ± 12.7, $P < 0.001$). This improvement was mainly attributable to an improved depiction of juxtacortical lesions (12.3 ± 10.8 vs 7.2 ± 5.6, $P < 0.001$). Interrater reliability was excellent in FLAIR 0.92 (95% confidence interval [CI], 0.85–0.95), synthDIR 0.93 (95% CI, 0.87–0.96), and trueDIR 0.95 (95% CI, 0.85–0.98).

Contrast-to-noise ratio in synthDIR exceeded that of FLAIR (22.0 ± 6.4 vs 16.7 ± 3.6, $P = 0.009$); no significant difference was seen in comparison to trueDIR (22.0 ± 6.4 vs 22.4 ± 7.9, $P = 0.87$).

Conclusions: Computationally generated DIR images improve lesion depiction compared with the use of standard modalities. This method demonstrates how artificial intelligence can help improving imaging in specific pathologies.

Key Words: artificial intelligence, deep learning, generative adversarial networks, multiple sclerosis, double inversion recovery

(*Invest Radiol* 2020;55: 318–323)

Multiple sclerosis (MS) is the most common nontraumatic cause of disability in young adults with accelerating rates of new diagnoses over the past decades.[1] Recently, data have accumulated demonstrating neuronal loss and early-onset atrophic brain changes in patients suffering from MS.[2,3] As the regenerative potential of brain tissue is per se limited and decreases with age, it is important to attenuate disease activity as early as possible in order to avoid nonrepairable damage. Both delayed initial diagnosis as well as an undetected progression under treatment can negatively impact the prognosis of patients.[4,5] Magnetic resonance imaging has proven to be a valuable tool in detecting and evaluating progression of MS-related brain lesions and is tightly integrated into current guidelines.[6] Although fluid-attenuated inversion recovery (FLAIR) is arguably the most widespread sequence used to assess brain lesions, technical advances have led to the development of double inversion recovery (DIR) acquisition with a higher sensitivity for lesion detection than conventional or fluid-attenuated T2-weighted acquisitions.[7] As it has recently become clear that (juxta)cortical plaques play an important role in MS and appear in early stages of the disease, the particular strength of DIR to detect these kinds of lesions is becoming more and more relevant from a clinical point of view.[8–11] Beyond the better conspicuity of (juxta)cortical lesions, DIR facilitates automated detection of lesions compared with standard sequences and thus seems better suited for segmentation tasks.[12,13] However, the time-consuming acquisition of DIR has hindered the widespread implementation of this technique in daily routine. In view of this limitation, synthesizing DIR from routinely acquired MR sequences with the aid of deep-learning tools, such as recently proposed generative adversarial networks (GAN), seems promising.[14] Reliable image synthesis with the here-proposed *DiamondGAN* (generative adversarial network with a diamond-shaped topology) has already been validated from a technical perspective.[15]

➢ See editorial on pages 324–325

Received for publication September 16, 2019; and accepted for publication, after revision, October 29, 2019.

From the *Department of Diagnostic and Interventional Neuroradiology, Klinikum rechts der Isar, Technical University of Munich, Munich; †Image-Based Biomedical Modeling at the Institute for Advanced Studies, Technical University of Munich, Garching; ‡Department of Neurology, Klinikum Rechts der Isar, Technical University of Munich, Munich; and §NeuroImaging Center, Klinikum Rechts der Isar, Technical University of Munich, Munich, Germany.

Conflicts of interest and sources of funding: none declared.
T.F. and H.L. share first authorship.

Correspondence to: Tom Finck, MD, Department of Diagnostic and Interventional Neuroradiology, Klinikum Rechts der Isar, Technical University Munich, Ismaninger Str. 22, 81675 Munich, Germany. E-mail: tom.finck@tum.de.

ISSN: 0020-9996/20/5505–0318
DOI: 10.1097/RLI.0000000000000640

An important aspect of *DiamondGAN* is its ability to augment image information via multi-to-one mapping, that is, to use several input sequences (in this case T1, T2, and FLAIR) to generate one output modality (in this case DIR). This enables *DiamondGAN* to learn synergistic combinations of image information from multiple input sequences and thus improves the output. In a prior work describing the technical background of *DiamondGAN*, we have shown DIR images of healthy controls, either acquired or synthesized, to 14 neuroradiologists and asked for their visual evaluation of both modalities.[15] Blinded to the origin of these images, they were unable to differentiate if the shown modality was acquired (trueDIR) or synthetic (synthDIR).

Whether GANs only create realistic looking images or whether these images indeed carry a surplus of clinically relevant information is however yet to be shown. We therefore investigated the hypothesis that synthDIR augments background image information from FLAIR, T1, and T2 and thus improves lesion depiction in MS patients.

MATERIALS AND METHODS

Scan Acquisition

This study design was approved by the local institutional review board, and informed consent was obtained from all patients. MR datasets from 100 randomly selected patients with diagnosed MS were retrospectively collected from a prospectively collected observation cohort from August 2014 to March 2016 and included T1 (2:25 minutes acquisition time), T2 (3:17 minutes acquisition time), FLAIR (3:55 minutes acquisition time), and DIR (6:31 minutes acquisition time) sequences. Scans were performed on a 3-T scanner (Philips Achieva 3.0 T; Philips Healthcare, Eindhoven, the Netherlands). Identical parameters in all patients were chosen for T1 (repetition time = 9 milliseconds, echo time = 4.0 milliseconds, flip angle = 8 degrees, acquired in the

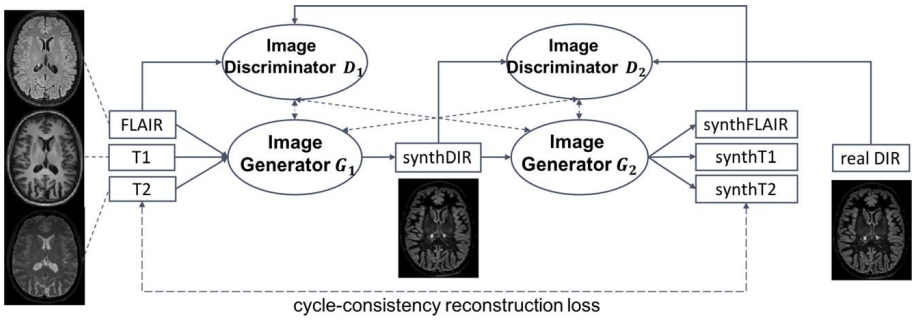


FIGURE 1. A schematic view of the proposed image synthesis system using generative adversarial networks, including 2 image generators and 2 image discriminators. The 4 networks are simultaneously optimized to generate high-quality images.

sagittal plane with isotropic voxel size of 1 mm^3), T2 (repetition time = 4000 milliseconds, echo time = 35.0 milliseconds, flip angle = 90 degrees, acquired in the sagittal plane with isotropic voxel size of 1 mm^3), FLAIR (repetition time = 10,000 milliseconds, echo time = 140.0 milliseconds, inversion time = 2750 milliseconds, flip angle = 90 degrees, acquired in the sagittal plane with isotropic voxel size of 1 mm^3), and DIR (repetition time = 5500 milliseconds, echo time = 321.6 milliseconds, inversion time = 2550 milliseconds and 2990 milliseconds, flip angle = 90 degrees, acquired in the sagittal plane with isotropic voxel size of 1 mm^3).

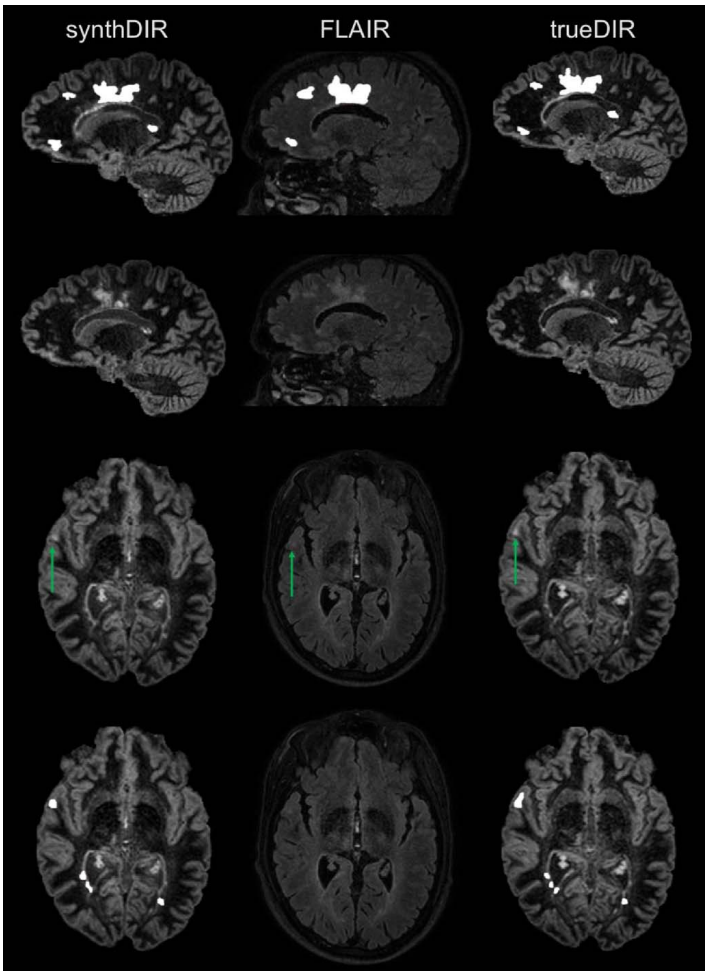


FIGURE 2. Exemplary images of FLAIR, synthDIR, and trueDIR from the same patient. Given are sets of slices in the sagittal and axial plane with their respective lesion segmentations. Notable is the improved ability to detect juxtacortical lesions (green arrows) in synthDIR compared with (input) FLAIR.

DIR Synthesis With GAN

The basic principle of *DiamondGAN* is to synthesize a target magnetic resonance imaging modality T given a set of *n* input modalities $X = \{x_i | i = 1, \ldots, n\}$. The goal of the synthesis task is to learn a modality generator G such that $G(X) = T$. The *DiamondGAN* pipeline contains 2 networks: a generator G and a discriminator D, built on conventional GAN techniques.[16] G contains 2 generators G_1 and G_2 that simultaneously learn the mappings from $X \rightarrow T$ and $T \rightarrow X$, respectively. D consists of 2 discriminators D_1 and D_2. D_1discriminates the real images from *source* domain and synthetic images from G_1, whereas D_2 discriminates the real images from *target* domain and synthetic images from G_2. In this adversarial learning process, the 4 networks are simultaneously optimized to generate high-quality images. G and D are variants of convolutional neural networks and optimized with multiple loss functions in an end-to-end fashion, as explained before.[15] One helpful property of publicly available *DiamondGAN* (https://github.com/hongweilibran/DiamondGAN/) is that it does not require the input and output to be strictly spatially aligned, by mapping the inputs to latent spaces and optimizing the generator networks by a cycle-consistency loss function as explained in Hongwei et al.[15] Conventional regression approaches require the input and output to be strictly spatially aligned. However, in practice, registration methods cannot guarantee such pixel-to-pixel alignment properly between the input and output image spaces.[17] A schematic overview illustrating the architecture of *DiamondGAN* is given in Figure 1.

We hypothesize that a large portion of information in an individual MR sequence is also contained (albeit possibly hidden) in other sequences and that DIR can be reconstructed given the combination of FLAIR, T1, and T2. The 3D volumes of these sequences are parsed into 2D axial slices. The concatenation of FLAIR, T1, and T2 axial slices is fed to train the network while synthDIR slices are the output. Technically, the network does not necessarily require 3D acquisitions for the sequences, but the acquisition should be consistent (either 2D or 3D)

TABLE 1. Patient Characteristics

	Training Set	Testing Set	*P*
n	50	50	1
Age	38.5 ± 9.2	36.4, 11.0	0.30
Sex	18 (36)	17 (34)	0.84
% RRMS	45 (90%)	46 (92%)	0.73
Disease duration	5.7 ± 4.6	5.8 ± 4.0	0.91
EDSS	1.5 (1.0–2.5)	1.5 (0–2.0)	0.61

Given are key clinical parameters for the training and testing set of the included 100 patients. RRMS indicates relapsing-remitting multiple sclerosis; EDSS, Expanded Disability Status Scale.

TABLE 2. Lesion Count (Mean ± Standard Deviation) as Well as ICCs (With 95% CI) for Both Readers and All Acquistions/Locations

	Periventricular			Juxtacortical		
	FLAIR	**synthDIR**	**trueDIR**	**FLAIR**	**synthDIR**	**trueDIR**
R1 counts	13.6 ± 9.35	16.7 ± 12.8	18.8 ± 14.3	7.2 ± 5.55	12.3 ± 10.8	14.7 ± 13.1
R2 counts	10.7 ± 8.36	13.0 ± 12.7	17.8 ± 15.1	3.7 ± 3.45	7.4 ± 9.6	11.1 ± 13.1
ICC	0.93 (0.88–0.96)	0.95 (0.91–0.97)	0.96 (0.93–0.98)	0.82 (0.68–0.90)	0.90 (0.82–0.94)	0.95 (0.91–0.97)
	Infratentorial			**Subcortical**		
	FLAIR	synthDIR	trueDIR	FLAIR	synthDIR	trueDIR
R1 counts	2.0 ± 1.68	2.4 ± 2.2	2.6 ± 2.2	12.1 ± 10.2	15.1 ± 10.8	15.8 ± 11.7
R2 counts	0.9 ± 1.26	1.3 ± 2.2	1.4 ± 1.6	14.6 ± 11.3	16.5 ± 12.2	18.8 ± 13.2
ICC	0.48 (0.07–0.70)	0.87 (0.78–0.93)	0.81 (0.66–0.89)	0.92 (0.87–0.96)	0.89 (0.81–0.94)	0.91 (0.83–0.95)

ICC indicates intraclass correlation coefficient; CI, confidence interval; FLAIR, fluid-attenuated inversion recovery; synthDIR, synthetic double inversion recovery; trueDIR, conventionally acquired DIR; R, reader.

among the input modalities. Fifty MS patients with complete sequences are used to train the model. The generator model contains 7,218,987 parameters and is trained for 12 hours with a high-end graphic card (NVIDIA Titan V, Santa Clara, CA). In the inference stage, we take FLAIR, T1, and T2 slices as the input and synthesize the DIR slices. Then the DIR slices are normalized by histogram matching and spatially concatenated into 3D volumes. Because the axial slices on top or bottom do not contain brain structures, we synthesize the middle slices with brain structure after empirically setting a starting and ending threshold. Exemplary sets of FLAIR, synthDIR, and trueDIR with their respective lesion segmentations are given in Figure 2.

Expert Readings

In accordance with the current imaging criteria to diagnose MS, we did not distinguish between juxtacortical and cortical lesions.[18] For simplicity, we will therefore refer to both types of lesions as juxtacortical lesions.

All 150 datasets (50 test patients with FLAIR, trueDIR, and synthDIR) were visually assessed by 2 neuroradiologists (both with 3 years of experience) for the number of juxtacortical, periventricular, infratentorial, and subcortical white matter lesions, having a minimum diameter of 3 mm in any direction. Manual lesion count was done independently using an open-source 3D image analysis tool,[19] and the order of investigated modalities was randomly altered to prevent a learning effect. The readers were blinded for the nature of DIR modalities (trueDIR or synthDIR) and had no clinical background information on patients except for the fact that prior diagnosis of MS had been made. Each lesion in synthDIR was retrospectively validated in trueDIR by one rater to exclude that false-positive lesions had been generated during the synthetization process.

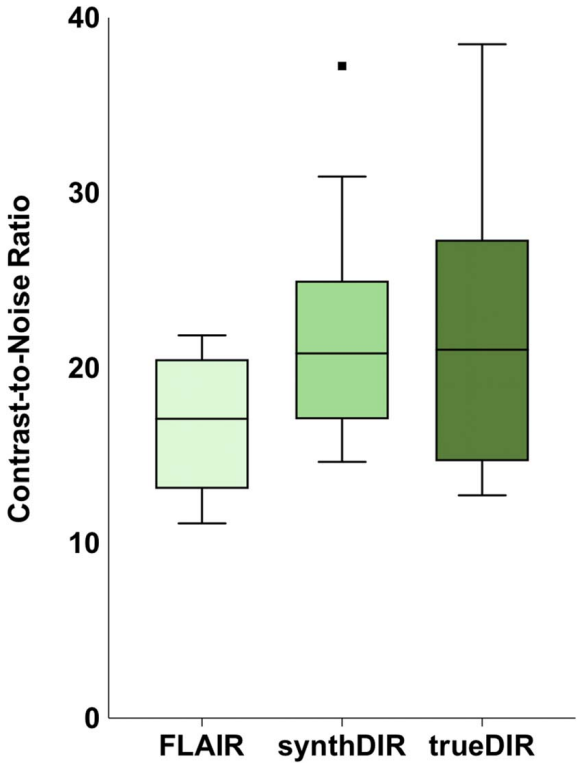


FIGURE 3. Contrast-to-noise ratios for FLAIR, synthDIR, and trueDIR.

Lesion Contrast

To assess image quality of the modalities, we calculated the contrast-to-noise ratio (CNR) for a randomly selected subset of 15 patients: in each patient, 1 or 2 representative lesions were manually segmented on the T2 image (to avoid bias), and equally sized regions of interest were placed in the contralateral normal-appearing white matter (NAWM) using open-source 3D image analysis tool.[19] From the coregistered modalities, CNR was calculated for FLAIR, synthDIR, and trueDIR as:

$$CNR = \frac{Mean_{SignalLesions} - Mean_{SignalNAWM}}{SD_{NAWM}}$$

Statistical Analysis

Same counts of lesions in FLAIR (that was used as input modality) and synthDIR was defined as null hypothesis. The normality of distribution was violated as tested by the D'Agostino-Pearson test. Lesion counts from the 3 investigated modalities were compared with a Wilcoxon signed-rank test; CNR between the investigated modalities was compared with a paired Student *t* test.

Interrater reliability was assessed with the intraclass correlation coefficient (ICC) (use of single measurements for absolute agreement in a 2-way random model).

Statistical computations were performed with software (SPSS Statistics for Windows, version 25.0; IBM, Armonk, NY). $P < 0.05$ was considered statistically significant.

RESULTS

Patient characteristics for the training and test set are given in Table 1. No significant differences in clinical parameters were observed.

Interrater reliability was excellent with intraclass correlation coefficients between both raters ranging from 0.92 (95% confidence interval [CI], 0.85–0.95) for FLAIR to comparable levels of 0.93 (95% CI, 0.87–0.96) and 0.95 (95% CI, 0.85–0.98) for synthDIR and trueDIR, respectively. Intraclass correlation coefficients between both readers as a function of lesion localization are given in Table 2.

Quantitative assessment of image contrast (lesion vs NAWM) in a subset of 15 randomly selected patients revealed a significantly better CNR for synthDIR compared with FLAIR (22.0 ± 6.4 vs 16.7 ± 3.6, $P = 0.009$), matching the CNR of trueDIR (22.0 ± 6.4 vs 22.4 ± 7.9, $P = 0.87$; Fig. 3).The mean count of MS-specific lesions (juxtacortical + periventricular + infratentorial) per patient was significantly higher in synthDIR than FLAIR (31.4 ± 20.7 vs 22.8 ± 12.7, $P < 0.001$; location-dependent lesion counts for both readers are given in Table 3; counts of MS-specific lesions are further shown in Fig. 4). Worth mentioning is the fact that the improved performance of synthDIR compared with FLAIR could be primarily attributed to better detection of juxtacortical lesions (12.3 ± 10.8 vs 7.2 ± 5.6, $P < 0.001$). Both observations held true when analyzing lesion counts from the second reader (Table 3).

Consequently, a shift in the proportion of juxtacortical and subcortical lesions was noted as the relative share of juxtacortical lesions from the total count increased from 16.9% in FLAIR to 23.3% in synthDIR ($P < 0.001$), whereas the share of subcortical lesions decreased analogously from 41.2% in FLAIR to 37.3% in synthDIR ($P < 0.001$).

Although a location-dependent heterogeneity could be noted, a tendency for improved lesion detection in synthDIR compared with FLAIR held true irrespective if lesions were juxtacortical, periventricular, subcortical, or infratentorial (counts for both readers are given in Table 3). Physically acquired trueDIR trumped both FLAIR (36.1 ± 24.3 vs 22.8 ± 12.7, $P < 0.001$) and synthDIR (36.1 ± 24.3 vs 31.4 ± 20.7, $P < 0.001$) in depicting MS-specific lesions. Retrospective visual cross-validation showed that there were no lesions in synthDIR that could not be detected in trueDIR; hence the possibility of artificial lesion generation by the GAN could be excluded.

DISCUSSION

We hypothesized that training a GAN with sequences routinely acquired in MS imaging allows to create synthetic, high lesion-to-background contrast DIR images. These generated images enabled the readers to find significantly more lesions compared with the standard sequence FLAIR.

Further supporting the ability of *DiamondGAN* to synergistically combine input image information to create its output is the fact that the increase in lesions found in the synthetic DIR compared with FLAIR was mostly driven by juxtacortical lesions. The DIR acquisition is best known for its ability to detect this type of lesion, as it has become increasingly clear that juxtacortical lesions play an important role for diagnosis and prognosis of MS patients.[20–22] By combining information from FLAIR, T1, and T2 images, *DiamondGAN* is able to replicate this strength of DIR images with potentially profound ramifications as initial diagnosis of MS strongly relies on the robust differentiation between MS-specific (ie, juxtacortical) non–MS-specific (ie, subcortical) lesions.

Beyond the methodological innovation of using complimentary aspects of different acquisitions (T1, T2, FLAIR) to generate output (synthDIR), such pronounced differences in lesion detection can have clinical implications for the monitoring of a disease that is tightly linked to the dynamics of inflammatory lesions. Recent studies have highlighted the potential of DIR in monitoring MS progression and stress the urgency to use this acquisition more broadly and prospectively wean our dependency on gadolinium scans.[22] In light of the time-consuming physical acquisition of DIR, artificial intelligence (AI) could be the key to facilitate its wider implementation in MS imaging. Even as the physically acquired DIR still outperformed the synthetic DIR, it needs to be said that the information content in synthDIR was nevertheless superior to that of input FLAIR and can potentially be improved with future improvements of GAN training and input combinations.

Beyond the subjective lesion analysis, quantification of CNR confirms the data augmentation taking place within *DiamondGAN*, a publicly available AI tool, and provides proof for the similitude between synthDIR and trueDIR in depicting white-matter lesions.

TABLE 3. Location-Dependent Lesion Count Differences, Discerned for Periventricular, Juxtacortical, Infratentorial, and Subcortical Lesions, as well as a Composite of All MS-Specific Lesions (Juxtacortical + Periventricular + Infratentorial)

	MS-Specific Lesions (PV + JC + IT)	*P*	Periventicular Lesions	*P*	Juxtacortical Lesions	*P*	Infratentorial Lesions	*P*	Subcortical Lesions	*P*
Reader 1										
FLAIR vs synthDIR	22.8 ± 12.7	<0.001*	13.6 ± 9.35	<0.001*	7.2 ± 5.55	<0.001*	2.0 ± 1.68	0.017*	12.1 ± 10.2	<0.001*
	31.4 ± 20.7		16.7 ± 12.8		12.3 ± 10.8		2.4 ± 2.2		15.1 ± 10.8	
FLAIR vs trueDIR	22.8 ± 12.7	<0.001*	13.6 ± 9.35	<0.001*	7.2 ± 5.55	<0.001*	2.0 ± 1.68	<0.001*	12.1 ± 10.2	<0.001*
	36.1 ± 24.3		18.8 ± 14.3		14.7 ± 13.1		2.6 ± 2.2		15.8 ± 11.7	
synthDIR vs trueDIR	31.4 ± 20.7	<0.001*	16.7 ± 12.8	0.002*	12.3 ± 10.8	0.004*	2.4 ± 2.2	0.14	15.1 ± 10.8	0.10
	36.1 ± 24.3		18.8 ± 14.3		14.7 ± 13.1		2.6 ± 2.2		15.8 ± 11.7	
Reader 2										
FLAIR vs synthDIR	15.3 ± 10.4	0.026*	10.7 ± 8.36	0.29	3.7 ± 3.45	0.0011*	0.86 ± 1.26	0.13	14.6 ± 11.3	0.074
	21.7 ± 20.8		13.0 ± 12.7		7.4 ± 9.6		1.3 ± 2.2		16.5 ± 12.2	
FLAIR vs trueDIR	15.3 ± 10.4	<0.001*	10.7 ± 8.36	<0.001*	3.7 ± 3.45	<0.001*	0.86 ± 1.26	0.0027*	14.6 ± 11.3	<0.001*
	30.3 ± 23.8		17.8 ± 15.1		11.1 ± 13.1		1.4 ± 1.6		18.8 ± 13.2	
synthDIR vs trueDIR	21.7 ± 20.8	<0.001*	13.0 ± 12.7	<0.001*	7.4 ± 9.6	<0.001*	1.3 ± 2.2	0.31	16.5 ± 12.2	0.0018*
	30.3 ± 23.8		17.8 ± 15.1		11.1 ± 13.1		1.4 ± 1.6		18.8 ± 13.2	

Counts are given for readers 1 and 2. *Significant results are highlighted.

MS indicates multiple sclerosis; PV, periventricular; JC, juxtacortical; IT, infratentorial; FLAIR, fluid-attenuated inversion recovery; synthDIR, synthetic double inversion recovery; trueDIR, conventionally acquired DIR.

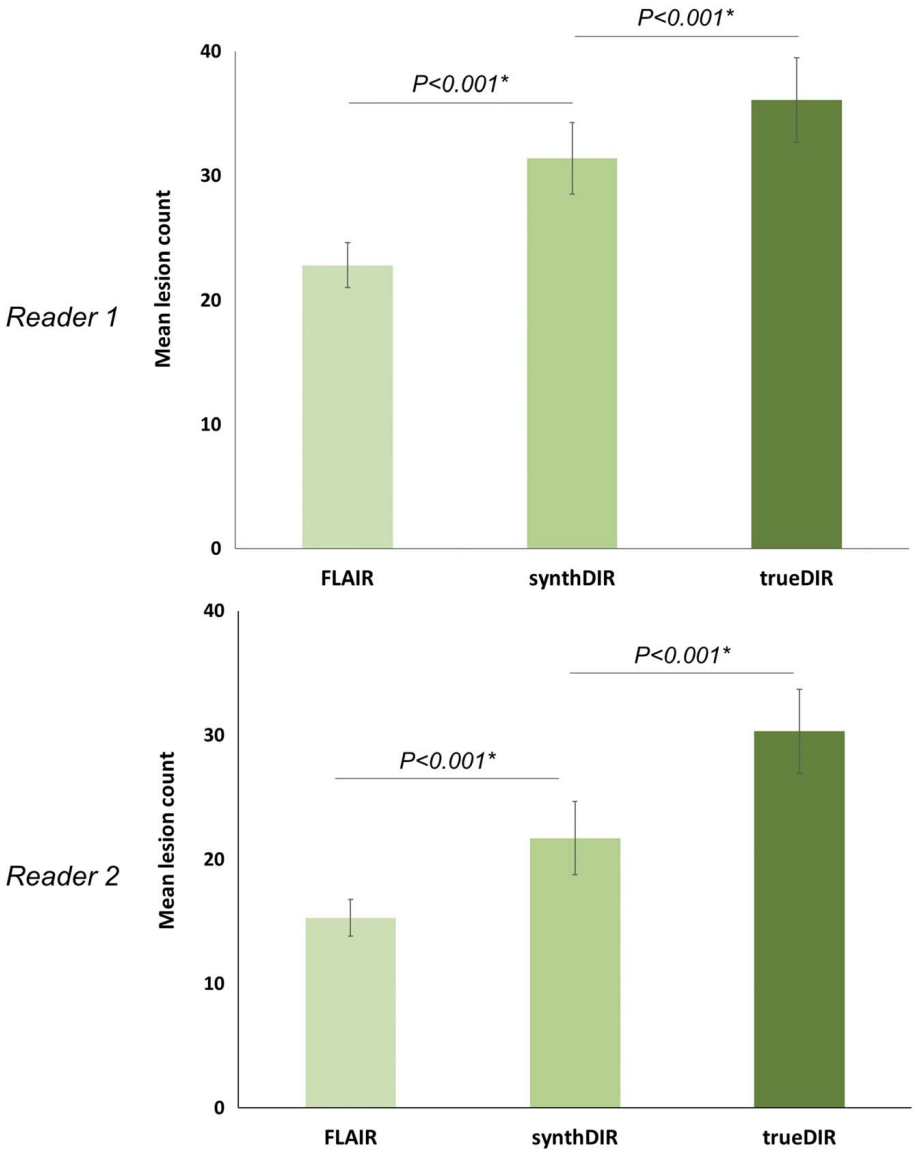


FIGURE 4. Mean counts of specific lesions (± standard error of the mean) for FLAIR, synthDIR, and trueDIR. Counts are given for both readers. Significant differences are highlighted with asterisk.

Although ongoing scientific advances, be it the improvement of image resolution, use of experimental methods such as MR-fingerprinting, or inclusion of spectroscopy findings, pave the way for improving diagnostics in MS, the utilization of AI in chronic inflammatory brain disease has in the majority of cases been restricted to quantitative analyses such as lesion segmentation.[23–26] The here-presented study provides a different approach as it does not focus on interpreting available images but explores the prospect of augmenting intrinsic, yet not necessarily visual information within an MR dataset. Similar approaches may further open the door for data homogenization through synthetization of a "standardized" MR dataset from heterogenous input data or generation of artificial imaging sets to feed deep-learning algorithms.

A further advantage of GANs is their applicability to existing data. This is in contrast to technical developments such as new sequences or hardware, which can only be applied prospectively and offers a unique chance to retrospectively validate findings that might become apparent after implementation of a more sensitive imaging protocol.[27]

One general limitation of this study is the relatively small sample size of only 100 patients. Further, the single-center setting with all scans originating from one MR scanner rendered statements about the generalizability impossible. Using GANs for data homogenization by synthesizing a standardized input dataset irrespective of the source data is however one potential remedy worth exploring. In addition, the input acquisitions used in our study (T1, T2, and FLAIR) all depict MS lesions, thus a large portion of information from the individual acquisition is redundant. Therefore, it remains to be studied whether satisfying synthetic images can be derived already from a subset of these acquisitions. Moreover, the value of alternative input acquisitions than the ones used in the present study has yet to be investigated.

In summary, we have demonstrated the ability of artificial neural networks to create high contrast images from standard input, thereby significantly improving lesion detection in MS patients. Future studies investigating generalizability and optimal sequence combinations for image synthesis seem warranted.

ACKNOWLEDGMENTS

M.M., M.B., and B.W. are supported by a DFG grant within the Priority Programme Radiomics.

REFERENCES

1. GBD 2016 Multiple Sclerosis Collaborators. Global, regional, and national burden of multiple sclerosis 1990–2016: a systematic analysis for the global burden of disease study 2016. *Lancet Neurol*. 2019;18:269–285.
2. Calabrese M, Atzori M, Bernardi V, et al. Cortical atrophy is relevant in multiple sclerosis at clinical onset. *J Neurol*. 2007;254:1212–1220.
3. Carassiti D, Altmann DR, Petrova N, et al. Neuronal loss, demyelination and volume change in the multiple sclerosis neocortex. *Neuropathol Appl Neurobiol*. 2018;44:377–390.
4. Rudick RA, Polman CH. Current approaches to the identification and management of breakthrough disease in patients with multiple sclerosis. *Lancet Neurol*. 2009;8:545–559.
5. Fernandez O, Fernandez V, Arbizu T, et al. Characteristics of multiple sclerosis at onset and delay of diagnosis and treatment in Spain (the novo study). *J Neurol*. 2010;257:1500–1507.
6. Filippi M, Rocca MA, Ciccarelli O, et al. MRI criteria for the diagnosis of multiple sclerosis: MAGNIMS consensus guidelines. *Lancet Neurol*. 2016;15:292–303.
7. Wattjes MP, Lutterbey GG, Gieseke J, et al. Double inversion recovery brain imaging at 3T: diagnostic value in the detection of multiple sclerosis lesions. *AJNR Am J Neuroradiol*. 2007;28:54–59.
8. Popescu BF, Bunyan RF, Parisi JE, et al. A case of multiple sclerosis presenting with inflammatory cortical demyelination. *Neurology*. 2011;76:1705–1710.
9. Redpath TW, Smith FW. Technical note: use of a double inversion recovery pulse sequence to image selectively grey or white brain matter. *Br J Radiol*. 1994;67:1258–1263.
10. Seewann A, Kooi EJ, Roosendaal SD, et al. Postmortem verification of MS cortical lesion detection with 3D DIR. *Neurology*. 2012;78:302–308.
11. Calabrese M, Filippi M, Gallo P. Cortical lesions in multiple sclerosis. *Nat Rev Neurol*. 2010;6:438–444.
12. Geurts JJ, Pouwels PJ, Uitdehaag BM, et al. Intracortical lesions in multiple sclerosis: improved detection with 3D double inversion-recovery MR imaging. *Radiology*. 2005;236:254–260.
13. Fartaria MJ, Bonnier G, Roche A, et al. Automated detection of white matter and cortical lesions in early stages of multiple sclerosis. *J Magn Reson Imaging*. 2016;43:1445–1454.
14. Nie D, Trullo R, Lian J, et al. Medical image synthesis with deep convolutional adversarial networks. *IEEE Trans Biomed Eng*. 2018;65:2720–2730.
15. Li H, Paetzold JC, Sekuboyina A, et al. *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019. MICCAI 2019. Lecture notes in computer science, vol 11767*. Cham: Springer; 2019.
16. Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. *Adv Neural Inf Process Syst*. 2014.
17. Alexander DC, Zikic D, Ghosh A, et al. Image quality transfer and applications in diffusion MRI. *Neuroimage*. 2017;152:283–298.
18. Thompson AJ, Banwell BL, Barkhof F, et al. Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. *Lancet Neurol*. 2018;17:162–173.
19. Yushkevich PA, Piven J, Hazlett HC, et al. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *Neuroimage*. 2006;31:1116–1128.
20. Moriarty DM, Blackshaw AJ, Talbot PR, et al. Memory dysfunction in multiple sclerosis corresponds to juxtacortical lesion load on fast fluid-attenuated inversion-recovery MR images. *AJNR Am J Neuroradiol*. 1999;20:1956–1962.
21. Stadelmann C, Albert M, Wegner C, et al. Cortical pathology in multiple sclerosis. *Curr Opin Neurol*. 2008;21:229–234.
22. Eichinger P, Schon S, Pongratz V, et al. Accuracy of unenhanced MRI in the detection of new brain lesions in multiple sclerosis. *Radiology*. 2019;291:429–435.
23. Llufriu S, Kornak J, Ratiney H, et al. Magnetic resonance spectroscopy markers of disease progression in multiple sclerosis. *JAMA Neurol*. 2014;71:840–847.
24. Granberg T, Uppman M, Hashim F, et al. Clinical feasibility of synthetic MRI in multiple sclerosis: a diagnostic and volumetric validation study. *AJNR Am J Neuroradiol*. 2016;37:1023–1029.
25. Commowick O, Istace A, Kain M, et al. Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure. *Sci Rep*. 2018;8:13650.
26. Kazuhiro K, Werner RA, Toriumi F, et al. Generative adversarial networks for the creation of realistic artificial brain magnetic resonance images. *Tomography*. 2018;4:159–163.
27. Eichinger P, Hock A, Schon S, et al. Acceleration of double inversion recovery sequences in multiple sclerosis with compressed sensing. *Invest Radiol*. 2019;54:319–324.

6 Lesion-specific, uncertainty-aware, and domain-adaptive image synthesis

This chapter has been published as one **peer-reviewed journal publication** and one **peer-reviewed conference publication**:

[1] T. Finck*, **H. Li***, S. Schlaeger, L. Grundl, N. Sollmann, B. Bender, E. Bürkle, C. Zimmer, J. Kirschke, B. Menze, et al. "Uncertainty-aware and lesion-specific image synthesis in multiple sclerosis magnetic resonance imaging: a multicentric validation study". In: ***Frontiers in Neuroscience*** 16 (2022)

[2] Q. Hu*, **H. Li***, and J. Zhang. "Domain-adaptive 3D medical image synthesis: an efficient unsupervised approach". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (**MICCAI**)* (2022), pp. 495–504

Synopsis: The above two works develop uncertainty-aware, lesion-specific and domain-adaptive neuroimage synthesis approaches. Specifically, publication ♯1 develops a novel generative adverarial network and a new loss function to provide uncertainty estimation and improve the synthesis quality of lesions. Furthermore, it is validated in a multi-center setting and demonstrates the effectiveness in MS lesion detection. Publication ♯2 explores domain adaptation for generative adversarial networks. First, it highlights the technical difference among the adaptation of image classification, image segmentation and image synthesis tasks. Second, it develops an efficient domain adaptation approach for 3D image synthesis based on 2D variational auto-encoder.

Contributions of thesis author: algorithm design and implementation, computational experiments and composition of manuscript.

frontiers | Frontiers in Neuroscience

ORIGINAL RESEARCH
published: 26 April 2022
doi: 10.3389/fnins.2022.889808

Uncertainty-Aware and Lesion-Specific Image Synthesis in Multiple Sclerosis Magnetic Resonance Imaging: A Multicentric Validation Study

Tom Finck[1]†, Hongwei Li[2†], Sarah Schlaeger[1], Lioba Grundl[1], Nico Sollmann[1,3], Benjamin Bender[4], Eva Bürkle[4], Claus Zimmer[1], Jan Kirschke[1], Björn Menze[2], Mark Mühlau[5,6] and Benedikt Wiestler[1,2]*

[1] Department of Diagnostic and Interventional Neuroradiology, School of Medicine, Klinikum rechts der Isar, Technical University of Munich, Munich, Germany, [2] Image-Based Biomedical Modeling, Technical University of Munich, Munich, Germany, [3] Department of Diagnostic and Interventional Radiology, University Hospital Ulm, Ulm, Germany, [4] Department of Diagnostic and Interventional Neuroradiology, Universitätsklinikum Tübingen, Tübingen, Germany, [5] TUM-Neuroimaging Center, Klinikum rechts der Isar, Technical University of Munich, Munich, Germany, [6] Department of Neurology, School of Medicine, Klinikum rechts der Isar, Technical University of Munich, Munich, Germany

OPEN ACCESS

Edited by:
Yogesh Rathi,
Harvard Medical School,
United States

Reviewed by:
Fan Zhang,
Harvard Medical School,
United States
Guoping Xu,
Wuhan Institute of Technology, China

***Correspondence:**
Tom Finck
tom.finck@tum.de

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: 04 March 2022
Accepted: 04 April 2022
Published: 26 April 2022

Citation:
Finck T, Li H, Schlaeger S, Grundl L, Sollmann N, Bender B, Bürkle E, Zimmer C, Kirschke J, Menze B, Mühlau M and Wiestler B (2022) Uncertainty-Aware and Lesion-Specific Image Synthesis in Multiple Sclerosis Magnetic Resonance Imaging: A Multicentric Validation Study. Front. Neurosci. 16:889808. doi: 10.3389/fnins.2022.889808

Generative adversarial networks (GANs) can synthesize high-contrast MRI from lower-contrast input. Targeted translation of parenchymal lesions in multiple sclerosis (MS), as well as visualization of model confidence further augment their utility, provided that the GAN generalizes reliably across different scanners. We here investigate the generalizability of a refined GAN for synthesizing high-contrast double inversion recovery (DIR) images and propose the use of uncertainty maps to further enhance its clinical utility and trustworthiness. A GAN was trained to synthesize DIR from input fluid-attenuated inversion recovery (FLAIR) and T1w of 50 MS patients (training data). In another 50 patients (test data), two blinded readers (R1 and R2) independently quantified lesions in synthetic DIR (synthDIR), acquired DIR (trueDIR) and FLAIR. Of the 50 test patients, 20 were acquired on the same scanner as training data (internal data), while 30 were scanned at different scanners with heterogeneous field strengths and protocols (external data). Lesion-to-Background ratios (LBR) for MS-lesions vs. normal appearing white matter, as well as image quality parameters were calculated. Uncertainty maps were generated to visualize model confidence. Significantly more MS-specific lesions were found in synthDIR compared to FLAIR (R1: 26.7 ± 2.6 vs. 22.5 ± 2.2 $p < 0.0001$; R2: 22.8 ± 2.2 vs. 19.9 ± 2.0, $p = 0.0005$). While trueDIR remained superior to synthDIR in R1 [28.6 ± 2.9 vs. 26.7 ± 2.6 ($p = 0.0021$)], both sequences showed comparable lesion conspicuity in R2 [23.3 ± 2.4 vs. 22.8 ± 2.2 ($p = 0.98$)]. Importantly, improvements in lesion counts were similar in internal and external data. Measurements of LBR confirmed that lesion-focused GAN training significantly improved lesion conspicuity. The use of uncertainty maps furthermore

helped discriminate between MS lesions and artifacts. In conclusion, this multicentric study confirms the external validity of a lesion-focused Deep-Learning tool aimed at MS imaging. When implemented, uncertainty maps are promising to increase the trustworthiness of synthetic MRI.

Keywords: magnetic resonance imaging, neuroradiology, multiple sclerosis, deep learning – artificial neural network (DL-ANN), double inversion recovery (DIR), synthetic MRI, artificial intelligence (AI)

INTRODUCTION

Magnetic resonance imaging (MRI) plays a central role in the management of patients with multiple sclerosis (MS), a neuroinflammatory disease with rising incidence that remains the most common cause of non-traumatic disability in the young (GBD 2016 Multiple Sclerosis Collaborators, 2019). MRI techniques have been developed to detect specific aspects of MS pathophysiology; double inversion recovery (DIR) imaging is exemplary of a sequence that improves lesion detection, in particular within the juxtacortical region. Through numerous studies, the superiority of DIR compared to established MRI sequences such as T2w or fluid-attenuated inversion recovery (FLAIR) sequences in depicting inflammatory white matter lesions has been validated (Geurts et al., 2005; Wattjes et al., 2007). Lengthy acquisition times and high technical requirements have, however, hindered the widespread use of DIR.

Recently, it has been shown that synthesizing DIR images with generative adversarial networks (GANs), a deep learning (DL) architecture with great potential for image synthesis, is feasible and improves lesion detection compared to FLAIR and T2w sequences (Finck et al., 2020; Bouman et al., 2021). Nonetheless, and in particular as MS lesions typically are small, GANs are at risk to synthesize images of high morphologic similarity to the target image, while failing to translate the clinically important MS lesions. Domain knowledge, i.e., the ability of a GAN to learn about the pathology-specific anomalies it should map, might open the door for further customization and improvements in this regard. Various classification tasks, from the categorization of breast lesions to the detection of malignant thyroid nodules have thus already been improved by complementing a network's training stage with domain knowledge (Feng et al., 2020; Avola et al., 2021). The underlying study is to our knowledge the first to investigate this knowledge-driven GAN approach in MS imaging.

The value of machine learning (ML) tools generally hinges on their ability to remain accurate when deployed to data that is of different structure from the training data, making multicentric validation a mandatory prerequisite. Also, building trust in artificial intelligence (AI) is oftentimes hindered because the decision-making process is concealed to the user who can only accept or discard a binary output (Asan et al., 2020). Hence, providing visibility into how an ML system makes predictions has become a major concern, especially in the medical domain (Quinn et al., 2022). This can be achieved either by providing insights into the "black-box" problem of DL systems that are inherently uninterpretable by the human operator or by designing networks that are inherently interpretable but generally less potent (i.e., linear regression, decision-trees). Neural networks are a hallmark of the "black-box" problem as decisions are made through nonlinear associations between input and output, thus remaining opaque to the human reader. Improved interpretability can be achieved by decreasing the complexity of such networks (i.e., reducing the amount of neural connections), at the potential cost of performance loss, or through uncertainty measurements of the decision-making process (Le et al., 2020). By providing uncertainty maps that quantify the decision-making confidence of a GAN, the acceptance of synthetic MRI by the medical community might be improved while also offering clearer insights into potential causes for a system's malfunctioning. Uncertainty maps can be estimated by analysis of the variance across iterations during image synthesis, which has of late become an area of increasing interest (Gal and Ghahramani, 2015; Watson et al., 2019). Visualization of model confidence in GAN-mediated synthesis of MRI has been done before in tasks such as artificial motion-artifact inclusion or age prediction in fetal MRI (Shaw et al., 2020; Shi et al., 2020). In contrast to these works, we aim to quantify model confidence in translating areas of pathology that only constitute a small fraction of the generated data volume.

This study presents a refined GAN framework with an architecture that includes a task-specific training objective for MS lesion translation. We hypothesize that this GAN-based approach can provide synthetic, high-contrast DIR images from routinely acquired input FLAIR and T1w data, thereby removing the need for time-intensive acquisition of DIR. A special focus of this study is to evaluate this task-specific network for external validity in a multicenter dataset with scanners from different vendors and different acquisition details. To further provide an insight into the decision-making process of the GAN and guide the reviewing clinician toward potential artifacts, we

Abbreviations: MRI, magnetic resonance imaging; MS, multiple sclerosis; DIR, double inversion recovery; FLAIR, fluid-attenuated inversion recovery; GAN, generative adversarial network; DL, deep learning; ML, machine learning; AI, artificial intelligence; synthDIR, synthetic double inversion recovery; trueDIR, physically acquired double inversion recovery; SSIM, structural similarity index measure; LST, lesion segmentation tool; JC, juxtacortical; PV, periventricular; IT, infratentorial; SC, subcortical; LBR, lesion-to-background ratios; LFL, lesion-focused loss; NAWM, normal appearing white matter; PSNR, peak signal-to-noise ratio; ICC, intraclass correlation coefficient.

calculated uncertainty maps that reflect the variance in image-to-image translation.

MATERIALS AND METHODS

Patients

The study design was approved by the local IRBs and informed consent was obtained from all patients at their respective centers prior to scan acquisition.

Training Data

Data for model training were retrospectively retrieved from 50 patients with diagnosed MS and included T1w (2:28 min), FLAIR (3:55 min), and DIR (6:31 min). All scans originated from the same scanner (Ingenia 3.0T, Philips Healthcare, Best, Netherlands). Sequence parameters were identical in all patients for T1w (TR of 9.0 ms, TE of 4.0 ms, flip angle of 8°, acquired in the sagittal plane with an isotropic voxel size of 1 mm^3), FLAIR (TR of 4,800 ms, TE of 331 ms, TI of 1,650 ms, flip angle of 90°, acquired in the sagittal plane with an isotropic voxel size of 1 mm^3), and DIR (TR of 5,500 ms, TE of 355.9 ms, TI of 2,550 ms and 2,990 ms, flip angle of 90°, acquired in the sagittal plane with an isotropic voxel size of 1.1 mm^3).

Testing Data

Sixty MRI scans from 50 consecutive patients (20:20:10 for centers 1:2:3, respectively) with diagnosed MS were included. For centers 1 and 2, 1 scan/patient was sampled, while baseline and follow-up exams for 10 patients from center 3 were considered. MRI data included T1w, FLAIR, and DIR and were acquired on both, 3.0T and 1.5T scanners. In detail, testing data from center 1 was acquired on the same hardware and using the same protocol as the training data (Ingenia 3.0T, Philips Healthcare, Best, Netherlands), testing data from center 2 originated from a different 3.0T scanner from the same manufacturer (Achieva 3.0T, Philips Healthcare, Best, Netherlands), and testing data from center 3 was acquired on 1.5T and 3.0T scanners from a different manufacturer (Skyra 3.0T, Avanto_fit 1.5T, and Aera 1.5T, Siemens Healthineers, Erlangen, Germany).

Sequence parameters for T1w, FLAIR, and DIR sequences were chosen according to the site-specific parameters optimized for routine clinical imaging and not modified during the retrieval period (**Supplementary Table 1**). Dichotomization of data from centers 1–3 was made to acknowledge the fact that data structure from (1) corresponded to the training data (prospectively referred to as "internal data"), while the data structure from (2) and (3) was unknown to the network (prospectively referred to as "external data"). **Table 1** illustrates how the data was categorized for evaluation.

TABLE 1 | Data from center 1 was acquired on the same hardware as training data and thus considered to be of known structure (= internal data).

Data class (number of image sets)	Classes for study evaluation
Training data (n = 50)	
Test data from (1) (n = 20)	Internal data (Known data structure)
Test data from (2) (n = 20)	External data (Unknown data structure)
Test data from (3) (n = 20)	External data (Unknown data structure)

In analogy, data from centers 2 and 3 were acquired on different hardware and considered to be of unknown structure (= external data).

Double Inversion Recovery Image Synthesis

Network Architecture

Our GAN extends the existing "pix2pix" method (Isola et al., 2017) and is trained to synthesize a target image y (resembling the true target image Y) given a set of input images X and a lesion segmentation mask S. In this setting, two networks compete with each other: The generator G is based on a U-Net architecture and synthesizes the target DIR images (synthDIR) from two input images (T1w and FLAIR), while the discriminator D tries to determine if a given DIR image is synthetic (synthDIR) or physically acquired (trueDIR). The network architecture and training process of the GAN are given in **Figure 1**. Importantly, the input of T1 and FLAIR images are fed to U-Net to generate DIR images while the lesion mask is only used to compute additional lesion-specific loss during the training stage (see below). Thus the lesion segmentation mask S is not required during inference.

Loss Functions

The discriminator gives the judgment about how realistic the local structures are (called "Patch GAN"), and is patch-based and driven by a least-square error (L2) loss function (Mao et al., 2019). The generator is trained on a composite loss function based on (a) the reconstruction error between the synthesized image and the target image using SSIM and (b) the output of the discriminator when judging if a given image is either ground truth or synthetic. In addition to an SSIM, a peculiarity of our model is that an additional loss focusing on the successful translation of MS lesions was developed. In order to focus the model on MS lesions (which only make up a minority of voxels in an image), an additional L1 loss term is calculated between the true and synthetic DIR images after multiplying both images with the lesion segmentation mask S, thus only considering the translation of MS lesions for this part of the loss. The image reconstruction loss for the generator G, the loss function for the discriminator D, and the total loss function were formulated as follows, respectively:

$$\mathcal{L}_{recons} = 1 - SSIM\,(Y,\ G\,(X)) + \lambda_1 * \ ||(Y - G(X)) \odot S_1|| \tag{1}$$

$$\mathcal{L}_D = \mathbb{E}_X\,\{||1 - D(X)||_2\} \tag{2}$$

$$\mathcal{L}_{total} = \lambda_2 * \mathcal{L}_{recons} + \mathcal{L}_D \tag{3}$$

Here, λ_1 and λ_2 are hyper-parameters and set to 1 and 10, respectively, which balances the two loss components.

Optimization

The input and output images were co-registered, skull-stripped, linearly transformed into the MNI152 space, and resampled

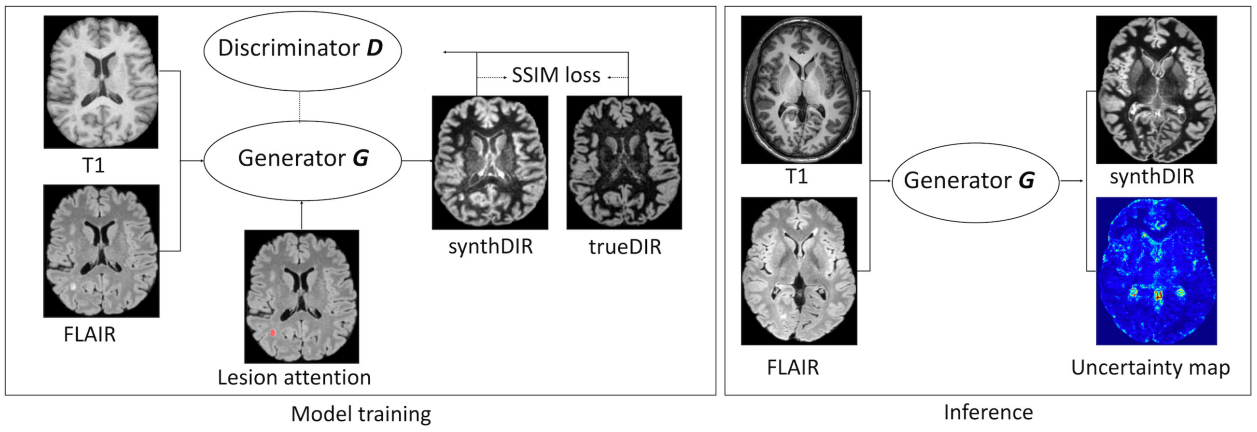


FIGURE 1 | Architecture, training process, and inference of the image synthesis task. The image Generator G uses the combination of FLAIR and T1w as input to generate synthDIR. The additional supervision from the lesion maps in the training stage drives an enhanced translation of MS-specific lesions (lesion attention). The feedback on the similarity between synthDIR and trueDIR is given by the Discriminator D and a structure similarity loss function and it updates the network weights until the loss function to discern both image pairs is minimal. During the inference stage, the trained generator G can generate the synthDIR and an uncertainty map showing the confidence of the output relating to each voxel. Uncertainty maps are calculated from the voxel-wise variances in signal intensities, as explained in the section "Materials and Methods".

to 1 mm isotropic resolution. As excellent correlation between automated and manual segmentation performance has been shown before, lesion segmentation maps were created using the Lesion Segmentation Tool (LST) (Schmidt et al., 2012). By including domain knowledge (in the form of lesion segmentation on FLAIR images) into the image translation during training, we enforced the model to pay attention to the lesion area by minimizing the difference between ground-truth images and synthetic images. In practice, such segmentation maps can be also provided by manual segmentation or other automated lesion segmentation tools (Schmidt et al., 2012; Li et al., 2018). Exemplary cases of all investigated sequences are shown in **Figure 2**. Training was carried out with a batch size of 1 for a total of 150 epochs, using the Adam optimizer with a learning rate of 0.001. During training, random intensity (gamma correction and gaussian blurring) and spatial (shifting and flipping) augmentations were performed. The best-performing model was selected using an internal validation set consisting of 10% of the training images.

The generated model is publicly available at https://figshare.com/articles/software/synthDIR/16607831.

Expert Readings

A dataset of 180 scans, comprising 60 sets each for FLAIR, synthDIR, and trueDIR, was investigated for lesion counts by two neuroradiologists (R1 with 5 years of experience in neuroradiological imaging, R2 with 3 years of experience in neuroradiological imaging) in a random order. Readers were blinded to scanner types and sequence labels. The number of juxtacortical (JC), periventricular (PV), infratentorial (IT), and subcortical (SC) lesions, in accordance with the 2017 McDonald criteria, were counted (Thompson et al., 2018). JC, PV, and IT lesions were considered to be MS-specific (Thompson et al., 2018). Albeit known to constitute different pathophysiological entities, we did not differentiate between cortical and juxtacortical lesions as this approach best reflects current guidelines (Bo et al., 2003; Thompson et al., 2018).

Quantitative Lesion Analysis and Uncertainty Maps

To quantitatively assess lesion translation, we calculated lesion-to-background ratios (LBR). Therefore, lesions on FLAIR and T1w images were segmented using LST, and tissue segmentation of T1w images was performed using ANTs Atropos (Avants et al., 2011). For comparison of LBR, GAN iterations with and without the above-stated lesion-specific loss function were computed.

From the segmentation maps, the lesion-to-background ratio was calculated as:

$$LBR = \frac{MeanSignal_{lesion}}{MeanSignal_{NAWM}} \quad (4)$$

Here, NAWM refers to "normal appearing white matter," i.e., non-lesioned white matter. From lesion segmentation maps and corresponding annotations in the NAWM, the mean signal intensity was extracted from DIR, FLAIR, and synthDIR images.

To estimate the GAN's uncertainty in generating synthDIR images, we performed variational inference during the test time by using dropout sampling. We added a dropout layer (dropout rate of 0.3) to the second-last layer of the U-Net and calculated 100 synthDIR images per input (Gal and Ghahramani, 2015). From these 100 iterations, we calculated the variance of voxel-wise intensities, resulting in the uncertainty map for visual inspection.

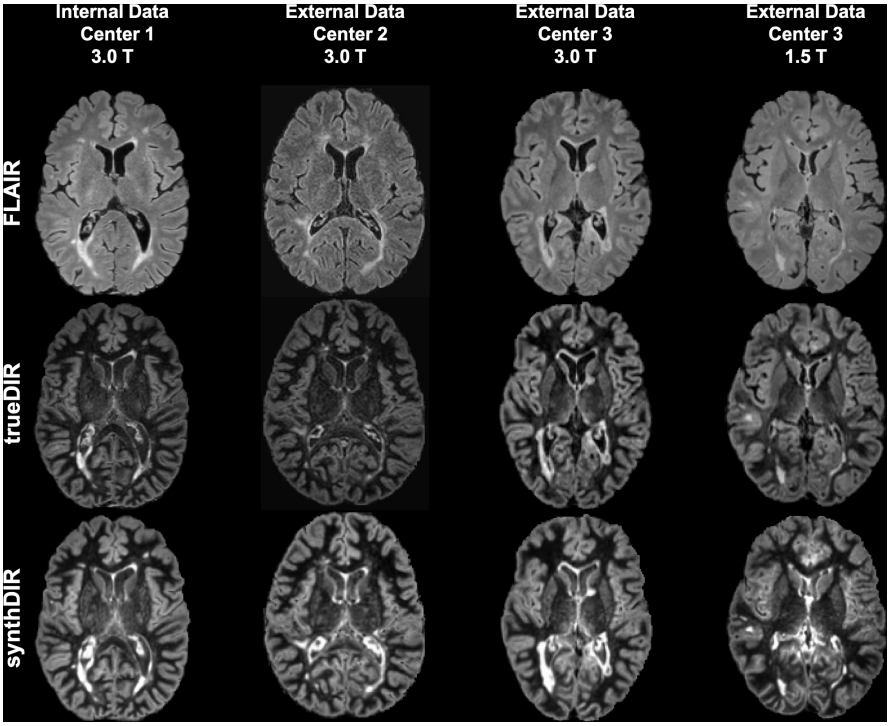


FIGURE 2 | Exemplary images of FLAIR, trueDIR, and synthDIR for all centers and scanners.

Statistical Analysis

Lesion counts were compared with a Wilcoxon signed-rank test to account for non-Gaussian distribution and paired data. LBR was compared with a paired *t*-test. Similarity of synthDIR and trueDIR was furthermore quantitatively assessed by the SSIM (Wang et al., 2004). For pixelwise comparisons, peak signal-to-noise ratio (PSNR) was calculated. Interrater agreement was assessed with the intraclass correlation coefficient (ICC; use of single measurements for absolute agreement in a two-way random model) and the related 95% confidence interval (95% CI). Statistical computations were performed with SPSS software (SPSS Statistics for Windows, version 25.0; IBM, Armonk, NY, United States). A p-value < 0.05 was considered statistically significant.

RESULTS

Interrater Agreement

Consistency between both readers was excellent with ICCs for all specific (JC + PV + IT) lesions ranging from 0.91 (95% CI: 0.85; 0.94) in FLAIR to 0.90 (95% CI: 0.84; 0.94) in synthDIR and 0.89 (95% CI: 0.83; 0.94) in trueDIR.

Lesion Counts

The study endpoint to improve depiction of MS specific lesions in synthDIR compared to FLAIR was met by both readers [26.7 ± 2.6 vs. 22.5 ± 2.2 ($p < 0.0001$) in R1 and 22.8 ± 2.2 vs. 19.9 ± 2.0 ($p = 0.0005$) in R2]. TrueDIR outperformed FLAIR in counts of MS-specific lesions [28.6 ± 2.9 vs. 22.5 ± 2.2 ($p < 0.0001$) in R1 and 23.3 ± 2.4 vs. 19.9 ± 2.0 ($p < 0.0001$) in R2]. While trueDIR remained superior to synthDIR in the depiction of MS-specific lesions in R1 [28.6 ± 2.9 vs. 26.7 ± 2.6 ($p = 0.0021$)], both image sets were of comparable diagnostic value in R2 [23.3 ± 2.4 vs. 22.8 ± 2.2 ($p = 0.98$)]. **Table 2** provides details on total and region-specific lesion counts for the study cohort.

Analysis of lesion counts as a function of scanner types revealed comparable effects independent of the structure of input data (internal or external). Hence, significant improvements in lesion counts were noted in synthDIR vs. FLAIR for both readers in external data [27.1 ± 3.4 vs. 22.6 ± 2.8 ($p < 0.0001$) in R1; 25.1 ± 2.9 vs. 21.5 ± 2.6 ($p = 0.0007$) in R2] and for R1 in internal data [26.6 ± 4.3 vs. 22.2 ± 3.6 ($p = 0.0029$) in R1; 18.1 ± 2.6 vs. 16.6 ± 2.6 ($p = 0.27$) in R2]. In external data, a slight improvement in lesion conspicuity was noted in trueDIR vs. synthDIR for R1 [28.9 ± 3.7 vs. 27.1 ± 3.4 ($p = 0.011$)] but not for R2 [25.6 ± 3.3 vs. 25.1 ± 2.9 ($p = 0.90$)]. **Table 3** provides lesion counts as a function of data source.

To increase the clinical reliability of synthDIR images, voxel-wise uncertainty maps from 100 forward runs using test-time dropout for Bayesian approximation were evaluated. For the majority of lesions, a high model confidence was observed, i.e., lesions were not highlighted in the uncertainty maps. On the other hand, artificial hyperintensities in synthetic images were readily identified by the high model uncertainty on

TABLE 2 | Lesion counts for all locations and both readers.

	All specific	*P*	PV lesions	*P*	JC lesions	*P*	IT lesions	*P*	SC lesions	*P*
Reader 1										
FLAIR vs. synthDIR	22.5 ± 2.2 vs. 26.7 ± 2.6	< 0.0001	12.0 ± 1.2 vs. 13.9 ± 1.4	< 0.0001	8.7 ± 1.2 vs. 10.8 ± 1.5	< 0.0001	1.9 ± 0.4 vs. 2.2 ± 0.4	0.043	10.6 ± 1.3 vs. 10.4 ± 1.2	0.82
FLAIR vs. trueDIR	22.5 ± 2.2 vs. 28.6 ± 2.9	< 0.0001	12.0 ± 1.2 vs. 13.9 ± 1.4	< 0.0001	8.7 ± 1.2 vs. 12.3 ± 1.7	< 0.0001	1.9 ± 0.4 vs. 2.4 ± 0.4	0.0002	10.6 ± 1.3 vs. 10.9 ± 1.4	0.36
SynthDIR vs. trueDIR	26.7 ± 2.6 vs. 28.6 ± 2.9	0.0021	13.9 ± 1.4 vs. 13.9 ± 1.4	0.91	10.8 ± 1.5 vs. 12.3 ± 1.7	< 0.0001	2.2 ± 0.4 vs. 2.4 ± 0.4	0.33	10.4 ± 1.2 vs. 10.9 ± 1.4	0.66
Reader 2										
FLAIR vs. synthDIR	19.9 ± 2.0 vs. 22.8 ± 2.2	0.0005	10.5 ± 1.0 vs. 12.4 ± 1.1	0.0004	7.8 ± 1.2 vs. 8.5 ± 1.3	0.18	1.5 ± 0.3 vs. 1.9 ± 0.3	0.024	13.5 ± 1.9 vs. 10.5 ± 1.5	< 0.0001
FLAIR vs. trueDIR	19.9 ± 2.0 vs. 23.3 ± 2.4	< 0.0001	10.5 ± 1.0 vs. 12.2 ± 1.2	0.0014	7.8 ± 1.2 vs. 9.7 ± 1.5	0.0028	1.5 ± 0.3 vs. 1.5 ± 0.3	0.99	13.5 ± 1.9 vs. 10.5 ± 1.6	< 0.0001
SynthDIR vs. trueDIR	22.8 ± 2.2 vs. 23.3 ± 2.4	0.98	12.4 ± 1.1 vs. 12.2 ± 1.2	0.26	8.5 ± 1.3 vs. 9.7 ± 1.5	0.068	1.9 ± 0.3 vs. 1.5 ± 0.3	0.03	10.5 ± 1.5 vs. 10.5 ± 1.6	0.70

PV, periventricular; JC, juxtacortical; IT, infratentorial; SC, subcortical; FLAIR, fluid-attenuated inversion recovery; trueDIR, real double inversion recovery; synthDIR, synthetic double inversion recovery.

TABLE 3 | Counts of MS-specific lesions for FLAIR, trueDIR, and synthDIR as a function of data source.

	All	*P*	Internal data	*P*	External data	*P*
Reader 1						
FLAIR vs. synthDIR	22.5 ± 2.2 vs. 26.7 ± 2.6	< 0.0001	22.2 ± 3.6 vs. 26.6 ± 4.3	0.0029	22.6 ± 2.8 vs. 27.1 ± 3.4	< 0.0001
FLAIR vs. trueDIR	22.5 ± 2.2 vs. 28.6 ± 2.9	< 0.0001	22.2 ± 3.6 vs. 27.9 ± 4.6	0.0001	22.6 ± 2.8 vs. 28.9 ± 3.7	< 0.0001
SynthDIR vs. trueDIR	26.7 ± 2.6 vs. 28.6 ± 2.9	0.0021	26.6 ± 4.3 vs. 27.9 ± 4.6	0.086	27.1 ± 3.4 vs. 28.9 ± 3.7	0.011
Reader 2						
FLAIR vs. synthDIR	19.9 ± 2.0 vs. 22.8 ± 2.2	0.0005	16.6 ± 2.6 vs. 18.1 ± 2.6	0.27	21.5 ± 2.6 vs. 25.1 ± 2.9	0.0007
FLAIR vs. trueDIR	19.9 ± 2.0 vs. 23.3 ± 2.4	< 0.0001	16.6 ± 2.6 vs. 18.6 ± 2.7	0.027	21.5 ± 2.6 vs. 25.6 ± 3.3	0.0001
SynthDIR vs. trueDIR	22.8 ± 2.2 vs. 23.3 ± 2.4	0.98	18.1 ± 2.6 vs. 18.6 ± 2.7	0.87	25.1 ± 2.9 vs. 25.6 ± 3.3	0.90

FLAIR, fluid-attenuated inversion recovery; trueDIR, real double inversion recovery; synthDIR, synthetic double inversion recovery.

these maps. **Figure 3** provides examples on how uncertainty maps allow to discern true-positive lesions from false-positive hyperintensities in synthDIR.

Quantitative Image Analysis

Similarity between trueDIR and synthDIR was highest in internal data, as shown by an SSIM of 0.967 ± 0.012, closely followed by external data (3) and (2) with still excellent SSIM-values of 0.950 ± 0.012 and 0.941 ± 0.010, respectively. For synthDIR, PSNR was highest in internal data at 29.2 ± 1.6 dB and decreased to 25.6 ± 1.1 dB in external data (3). **Table 4** provides detailed values for quantitative image metrics.

Effects of Lesion-Focused Loss Function

To assess the benefit of the lesion-specific loss function during image synthesis, LBR were compared between FLAIR, trueDIR, synthDIR, as well as synthDIR generated by a network iteration without the lesion-specific loss. Both versions of synthDIR, irrespective if additional loss was included or not, exceeded input FLAIR in LBR (data given in **Table 4**).

Of note, LBR was significantly lower in synthDIR generated by the version without lesion-focused loss compared to the version of synthDIR benefiting from lesion-focused loss (2.69 ± 0.66 vs. 2.80 ± 0.67, $p < 0.001$). While synthDIR achieved a comparable LBR to trueDIR (2.80 ± 0.67 vs. 2.86 ± 0.65, $p = 0.41$), this effect faded if synthDIR was generated without lesion-focused loss (2.69 ± 0.66 vs. 2.86 ± 0.65, $p = 0.032$) (as shown in **Figure 4**).

DISCUSSION

Medical imaging has benefited greatly from DL advances that gave birth to a panoply of systems aimed at tasks ranging from disease detection to image synthesis and artifact reduction (Emami et al., 2018; Rajpurkar et al., 2018; Liang et al., 2019). We here validated a GAN that has been fine-tuned to the translation of MS-specific white matter lesions while aiming to remain generalizable to external data. We further explored the concept of uncertainty maps to illustrate how trustworthy the network is in image-to-image translation. Such maps can

TABLE 4 | Image-wise (SSIM) and voxel-wise (PSNR) comparative metrics for synthDIR and trueDIR.

	SSIM (trueDIR – synthDIR)	PSNR (dB) (trueDIR – synthDIR)	LBR FLAIR	LBR trueDIR	LBR synthDIR	LBR synthDIR w/o LFL
All	0.954 ± 0.016	27.2 ± 2.2	1.52 ± 0.49	2.86 ± 0.65	2.80 ± 0.67	2.69 ± 0.66
Internal data	0.967 ± 0.012	29.2 ± 1.64	1.45 ± 0.06	2.80 ± 0.33	2.86 ± 0.34	2.68 ± 0.30
External data (2)	0.941 ± 0.010	25.8 ± 1.12	1.65 ± 0.12	3.01 ± 0.41	3.35 ± 0.50	3.31 ± 0.45
External data (3)	0.950 ± 0.012	25.6 ± 1.08	1.46 ± 0.86	2.78 ± 1.00	2.19 ± 0.56	2.07 ± 0.50

LBR are given for FLAIR, trueDIR, synthDIR, as well as for synthDIR generated by a GAN iteration without the lesion-focused loss function (synthDIR w/o LFL). Results are given for internal data, as well as external data (2) and (3). SSIM, structural similarity index measure; PSNR, peak signal-to-noise ratio; LBR, lesion-to-background ratio; LFL, lesion-focused loss; trueDIR, real double inversion recovery; synthDIR, synthetic double inversion recovery.

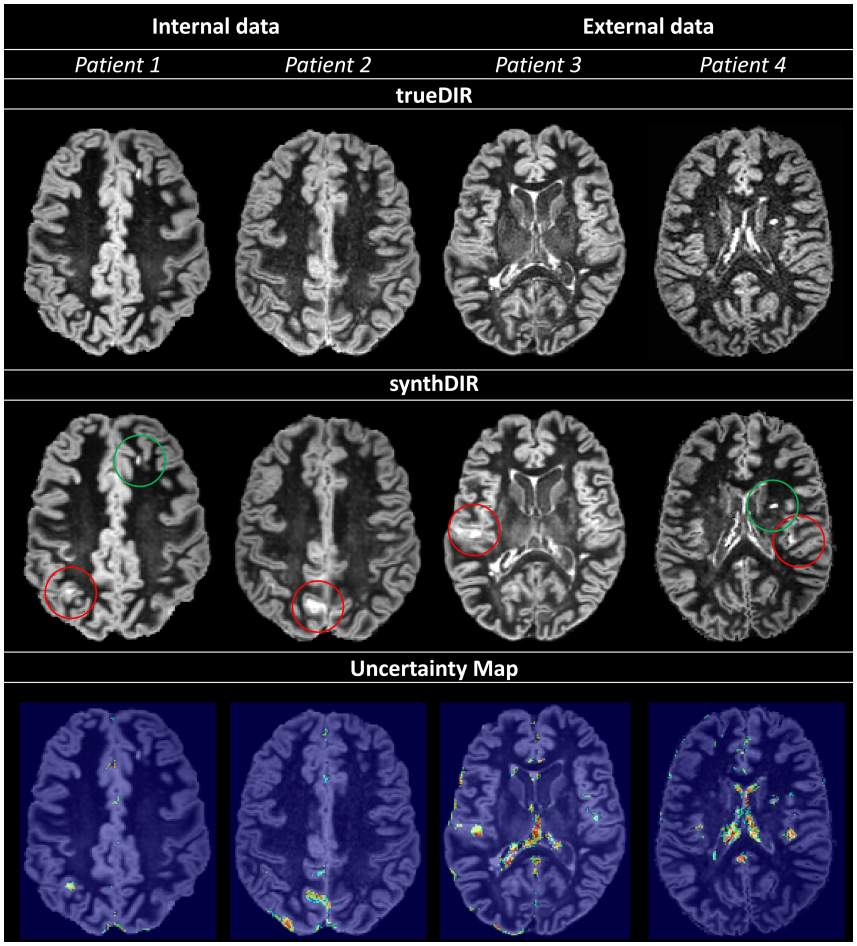


FIGURE 3 | Uncertainty maps provide relevant information regarding the validity of voxel-to-voxel translation; increases in uncertainty are scaled from blue to green. Circled in red (Patients 1–4) are hyperintensities in synthDIR without correlation in trueDIR and easily recognized as areas of high variance in the corresponding uncertainty maps, allowing for their identification as artifacts from the synthesis task. On the other hand, true-positive lesions are readily identified as regions with either no (patient 1 – green circle in synthDIR) or low (patient 4 – green circle in synthDIR) values of uncertainty. Hence, interpretation of synthDIR and decision-making on the veracity of lesions is facilitated through uncertainty maps.

provide important support to decide on the veracity of findings in synthetic images and help the radiologist to detect artifacts resulting from the synthesis task.

Comparison of the network's performance in internal and external data showed that significantly more MS-specific lesions could be found in synthDIR compared to the FLAIR sequence that was used as input, irrespective of the data origin. Approximately 20% more MS-specific lesions were thus depictable in synthDIR, a magnitude of difference that is of obvious clinical interest, especially in patients with low lesion counts. While other surrogates of MS activity have been explored, depiction of new inflammatory plaques is still considered the hallmark of disease monitoring in MS (Filippi et al., 2001; Chard et al., 2003; Wattjes et al., 2015). Also, lesion load has

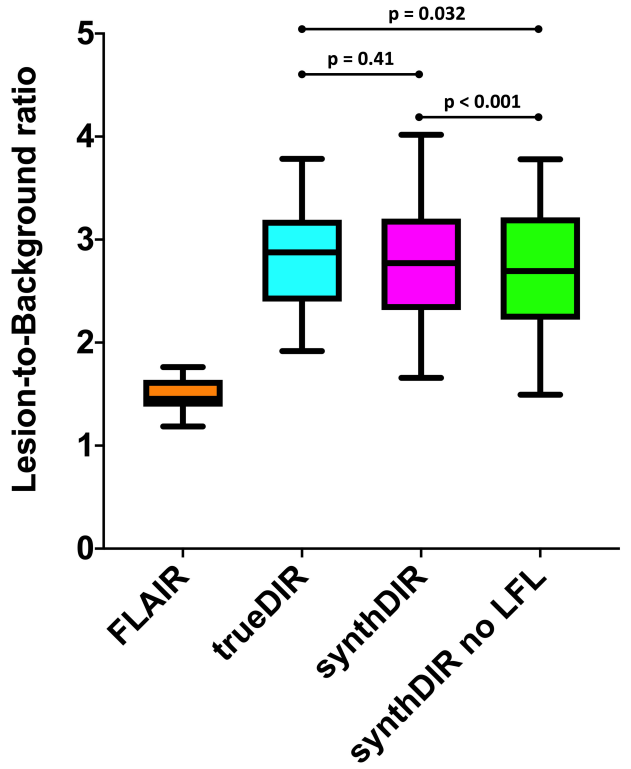


FIGURE 4 | Lesion-to-background ratios for FLAIR, trueDIR, and synthDIR. Additionally, LBR was calculated for synthDIR generated by a GAN-iteration without the lesion-focused loss (synthDIR no LFL). Of note, LBR was significantly higher in synthDIR compared to synthDIR without LFL, confirming the hypothesis that domain knowledge can be improved through LFL. While there was no significant difference in LBR between synthDIR and trueDIR (p = 0.41), LBR of synthDIR without LFL remained inferior to the LBR of trueDIR (p = 0.032). LBR, lesion-to-background ratio; LFL, lesion-focused loss.

been shown to directly correlate with future disability and, if properly detected and reliably quantified, might therefore prompt escalation of disease-modifying therapy (Calabrese et al., 2010; Popescu et al., 2011).

Domain knowledge, i.e., the ability to learn about pathology-specific image findings, is promising to further augment the clinical utility of DL tools (Yuan et al., 2019). The improved lesion translation that our GAN achieved by including a lesion-focused loss function hints at the potential of domain knowledge to further customize synthetic imaging. To highlight this, we showed that LBR in synthDIR was non-inferior to LBR in trueDIR only if the GAN was complemented by a lesion-focused loss.

The ability of synthDIR to outperform FLAIR, a sequence still considered gold-standard in MS imaging, has been shown for a multi-modal input (T1w, T2w, and FLAIR) in a monocentric setting (Finck et al., 2020; Bouman et al., 2021). In doing so, relevant reductions in scan times are feasible as the physical acquisition of 3D and isotropic DIR may take up to 7 min (Eichinger et al., 2019). While other methods, such as sparse sampling, have previously achieved scan time reductions for DIR, a GAN-based approach might be advantageous as it works on existing data and thus does not need to be prospectively deployed (Eichinger et al., 2019). This offers the potential advantage to augment the diagnostic value of existing studies and, hence, to render longitudinal follow-up exams more conclusive.

Albeit accurate in their output, neural networks generally fail to provide insight into the decision-making process, the so-called "black-box problem." Rendering this process more transparent is crucial for the acceptance of said networks and can, in theory, be achieved by providing methods to interpret the "black-box," or by designing models that are inherently more transparent in their functioning (Rudin, 2019; Arun et al., 2020). In GANs specifically, one potential bias in trying to match the (lesion) distribution in the target domain (trueDIR) is that features (lesions) with no correlation in source data might be erroneously mapped, a phenomenon commonly referred to as "hallucination." To verify lesion veracity we therefore introduced the concept of uncertainty maps that highlight the voxel-wise aleatoric variance taking place during image translation. Hence, the ability to compare hyperintensities in synthDIR to their respective uncertainty mappings can reduce the risk of false-positive findings, i.e., misinterpretation of constructed lesions in the synthetic image data. **Figure 3** illustrates how MS lesions can thus be separated from artifacts according to their voxel-wise intensity variance. As erroneous mappings remain an intrinsic limitation of GANs, their future deployment might benefit greatly from the calculation of uncertainty maps that are displayed in parallel to synthetic images.

A limitation of this approach is that having to reference synthDIR, along with the uncertainty maps adds complexity to the longitudinal interpretation of clinical MRI. Furthermore, comparison of synthDIR and trueDIR via autosegmentation techniques might have provided more objective lesion counts in this study. However, as our GAN was designed to provide synthetic data for clinical use, we opted for manual lesion counts as this best reflects the clinical reality. Future iterations of synthDIR might furthermore mitigate the wide disparities in lesion counts that we noticed especially in SC lesions. Also, prospective investigations should explore the feasibility to generate a GAN targeted to create synthDIR while using even fewer, potentially only one input modality. At last, we tested for generalizability by including three centers with differing hardware. Future investigations would benefit from the inclusion of more centers and readers, as our results demonstrate equivalence of synthDIR to trueDIR for only one of the two neuroradiologists.

CONCLUSION

Our findings confirm the use-case and external validity of a DL tool targeted at improving MRI in patients with MS. Our study demonstrates both, the utility of lesion-focused learning to improve domain adaption, as well as the potential benefit of uncertainty maps to help gain trust in GANs and make informed medical decisions. Presumably, wider deployment of these tools

could prove beneficial in MS where treatment decisions are heavily relying on MRI findings.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because of National Data Protection Law. Requests to access the datasets should be directed to TF, tom.finck@tum.de.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethikkommission Klinikum rechts der Isar. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

BW, HL, BM, CZ, JK, TF, LG, and MM conceived and designed the project. BB and EB conceived the study and contributed external datasets. HL, BM, and BW designed the GAN. NS and SS performed the experiments. TF and BW prepared the writing of the first draft and performed the statistical analyses. TF prepared the figures and tables. All authors reviewed the first draft of the manuscript, contributed to the article, and approved the final manuscript draft.

FUNDING

BM, MM, and BW were supported by the DFG, SPP Radiomics. HL was supported by the Forschungskredit (Grant No. FK-21-125) from the University of Zurich.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2022.889808/full#supplementary-material

Supplementary Figure 1 | Network architecture: **(A)** The U-Net generator used to produce a synthetic DIR image from FLAIR and T1 input images. **(B)** The patch-based discriminator which receives as input both the source image (T1 and FLAIR) and either a real or synthetic DIR. The discriminator is patch-based. ks, kernel size.

REFERENCES

Arun, N., Gaw, N., Singh, P., Chang, K., Aggarwal, M., Chen, B., et al. (2020). Assessing the (Un)Trustworthiness of saliency maps for localizing abnormalities in medical imaging. *arXiv* [Preprint]. arXiv:2008.02766 doi: 10.1148/ryai.2021200267

Asan, O., Bayrak, A. E., and Choudhury, A. (2020). Artificial intelligence and human trust in healthcare: focus on clinicians. *J. Med. Internet Res.* 22:e15154. doi: 10.2196/15154

Avants, B. B., Tustison, N. J., Wu, J., Cook, P. A., and Gee, J. C. (2011). An open source multivariate framework for n-tissue segmentation with evaluation on public data. *Neuroinformatics* 9, 381–400. doi: 10.1007/s12021-011-9109-y

Avola, D., Cinque, L., Fagioli, A., Filetti, S., Grani, G., and Rodolà, E. (2021). Multimodal feature fusion and knowledge-driven learning via experts consult for thyroid nodule classification. *IEEE Trans. Circuits Syst. Video Technol.* 1. doi: 10.1109/TCSVT.2021.3074414

Bo, L., Vedeler, C. A., Nyland, H. I., Trapp, B. D., and Mork, S. J. (2003). Subpial demyelination in the cerebral cortex of multiple sclerosis patients. *J. Neuropathol. Exp. Neurol.* 62, 723–732. doi: 10.1093/jnen/62.7.723

Bouman, P. M., Strijbis, V. I., Jonkman, L. E., Hulst, H. E., Geurts, J. J., and Steenwijk, M. D. (2021). Artificial double inversion recovery images for (juxta)cortical lesion visualization in multiple sclerosis. *Mult. Scler.* 28, 541–549. doi: 10.1177/1352458521102986O

Calabrese, M., Filippi, M., and Gallo, P. (2010). Cortical lesions in multiple sclerosis. *Nat. Rev. Neurol.* 6, 438–444. doi: 10.1038/nrneurol.2010.93

Chard, D. T., Brex, P. A., Ciccarelli, O., Griffin, C. M., Parker, G. J., Dalton, C., et al. (2003). The longitudinal relation between brain lesion load and atrophy in multiple sclerosis: a 14 year follow up study. *J. Neurol. Neurosurg. Psychiatry* 74, 1551–1554. doi: 10.1136/jnnp.74.11.1551

Eichinger, P., Hock, A., Schon, S., Preibisch, C., Kirschke, J. S., Muhlau, M., et al. (2019). Acceleration of double inversion recovery sequences in multiple sclerosis with compressed sensing. *Invest. Radiol.* 54, 319–324. doi: 10.1097/RLI.0000000000000550

Emami, H., Dong, M., Nejad-Davarani, S. P., and Glide-Hurst, C. K. (2018). Generating synthetic CTs from magnetic resonance images using generative adversarial networks. *Med. Phys.* 45, 3627–3636. doi: 10.1002/mp.13047

Feng, H., Cao, J., Wang, H., Xie, Y., Yang, D., Feng, J., et al. (2020). A knowledge-driven feature learning and integration method for breast cancer diagnosis on multi-sequence MRI. *Magn. Reson. Imaging* 69, 40–48. doi: 10.1016/j.mri.2020.03.001

Filippi, M., Cercignani, M., Inglese, M., Horsfield, M. A., and Comi, G. (2001). Diffusion tensor magnetic resonance imaging in multiple sclerosis. *Neurology* 56, 304–311. doi: 10.1212/wnl.56.3.304

Finck, T., Li, H., Grundl, L., Eichinger, P., Bussas, M., Muhlau, M., et al. (2020). Deep-learning generated synthetic double inversion recovery images improve multiple sclerosis lesion detection. *Invest. Radiol.* 55, 318–323. doi: 10.1097/RLI.0000000000000640

Gal, Y., and Ghahramani, Z. (2015). "Dropout as a Bayesian approximation: representing model uncertainty in deep learning," in *Proceedings of the 33rd International Conference on Machine Learning*, New York, NY. doi: 10.3390/s20216011

GBD 2016 Multiple Sclerosis Collaborators (2019). Global, regional, and national burden of multiple sclerosis 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet Neurol.* 18, 269–285. doi: 10.1016/S1474-4422(18)30443-5

Geurts, J. J., Pouwels, P. J., Uitdehaag, B. M., Polman, C. H., Barkhof, F., and Castelijns, J. A. (2005). Intracortical lesions in multiple sclerosis: improved detection with 3D double inversion-recovery MR imaging. *Radiology* 236, 254–260. doi: 10.1148/radiol.2361040450

Isola, P., Zhu, J., Zhou, T., and Efros, A. (2017). "Image-to-image translation with conditional adversarial networks," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI. doi: 10.1109/CVPR.2017.632

Le, V., Quinn, T. P., Tran, T., and Venkatesh, S. (2020). Deep in the bowel: highly interpretable neural encoder-decoder networks predict gut metabolites from gut microbiome. *BMC Genomics* 21(Suppl. 4):256. doi: 10.1186/s12864-020-6652-7

Li, H., Jiang, G., Zhang, J., Wang, R., Wang, Z., Zheng, W. S., et al. (2018). Fully convolutional network ensembles for white matter hyperintensities

segmentation in MR images. *Neuroimage* 183, 650–665. doi: 10.1016/j.neuroimage.2018.07.005

Liang, K., Zhang, L., Yang, H., Yang, Y., Chen, Z., and Xing, Y. (2019). Metal artifact reduction for practical dental computed tomography by improving interpolation-based reconstruction with deep learning. *Med. Phys.* 46, e823–e834. doi: 10.1002/mp.13644

Mao, X., Li, Q., Xie, H., Lau, R. Y. K., Wang, Z., and Smolley, S. P. (2019). On the effectiveness of least squares generative adversarial networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 41, 2947–2960. doi: 10.1109/TPAMI.2018.2872043

Popescu, B. F., Bunyan, R. F., Parisi, J. E., Ransohoff, R. M., and Lucchinetti, C. F. (2011). A case of multiple sclerosis presenting with inflammatory cortical demyelination. *Neurology* 76, 1705–1710. doi: 10.1212/WNL.0b013e31821a44f1

Quinn, T. P., Jacobs, S., Senadeera, M., Le, V., and Coghlan, S. (2022). The three ghosts of medical AI: can the black-box present deliver? *Artif. Intell. Med.* 124:102158. doi: 10.1016/j.artmed.2021.102158

Rajpurkar, P., Irvin, J., Ball, R. L., Zhu, K., Yang, B., Mehta, H., et al. (2018). Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Med.* 15:e1002686. doi: 10.1371/journal.pmed.1002686

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* 1, 206–215. doi: 10.1038/s42256-019-0048-x

Schmidt, P., Gaser, C., Arsic, M., Buck, D., Forschler, A., Berthele, A., et al. (2012). An automated tool for detection of FLAIR-hyperintense white-matter lesions in multiple sclerosis. *Neuroimage* 59, 3774–3783. doi: 10.1016/j.neuroimage.2011.11.032

Shaw, R., Sudre, C., Varsavsky, T., Ourselin, S., and Cardoso, M. J. (2020). A k-space model of movement artefacts: application to segmentation augmentation and artefact removal. *IEEE Trans. Med. Imaging* 39, 2881–2892. doi: 10.1109/TMI.2020.2972547

Shi, W., Yan, G., Li, Y., Li, H., Liu, T., Sun, C., et al. (2020). Fetal brain age estimation and anomaly detection using attention-based deep ensembles with uncertainty. *Neuroimage* 223:117316. doi: 10.1016/j.neuroimage.2020.117316

Thompson, A. J., Banwell, B. L., Barkhof, F., Carroll, W. M., Coetzee, T., Comi, G., et al. (2018). Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. *Lancet Neurol.* 17, 162–173. doi: 10.1016/S1474-4422(17)30470-2

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* 13, 600–612. doi: 10.1109/tip.2003.819861

Watson, D. S., Krutzinna, J., Bruce, I. N., Griffiths, C. E., McInnes, I. B., Barnes, M. R., et al. (2019). Clinical applications of machine learning algorithms: beyond the black box. *BMJ* 364:l886. doi: 10.1136/bmj.l886

Wattjes, M. P., Lutterbey, G. G., Gieseke, J., Traber, F., Klotz, L., Schmidt, S., et al. (2007). Double inversion recovery brain imaging at 3T: diagnostic value in the detection of multiple sclerosis lesions. *AJNR Am. J. Neuroradiol.* 28, 54–59.

Wattjes, M. P., Rovira, A., Miller, D., Yousry, T. A., Sormani, M. P., de Stefano, M. P., et al. (2015). Evidence-based guidelines: MAGNIMS consensus guidelines on the use of MRI in multiple sclerosis–establishing disease prognosis and monitoring patients. *Nat. Rev. Neurol.* 11, 597–606. doi: 10.1038/nrneurol.2015.157

Yuan, W., Wei, J., Wang, J., Ma, Q., and Tasdizen, T. (2019). Unified attentional generative adversarial network for brain tumor segmentation from multimodal unpaired images. *arXiv* [Preprint]. arXiv:1907.03548

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Domain-Adaptive 3D Medical Image Synthesis: An Efficient Unsupervised Approach

Qingqiao Hu[1], Hongwei Li[2,3], and Jianguo Zhang[1( )]

[1] Research Institute of Trustworthy Autonomous System, Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, China
zhangjg@sustech.edu.cn
[2] Department of Computer Science, Technical University of Munich, Munich, Germany
hongwei.li@tum.de
[3] Department of Quantitative Biomedicine, University of Zurich, Zürich, Switzerland

Abstract. Medical image synthesis has attracted increasing attention because it could generate missing image data, improve diagnosis, and benefits many downstream tasks. However, so far the developed synthesis model is not adaptive to unseen data distribution that presents domain shift, limiting its applicability in clinical routine. This work focuses on exploring domain adaptation (DA) of *3D* image-to-image synthesis models. First, we highlight the technical difference in DA between classification, segmentation, and synthesis models. Second, we present a novel efficient adaptation approach based on a 2D variational autoencoder which approximates 3D distributions. Third, we present empirical studies on the effect of the amount of adaptation data and the key hyperparameters. Our results show that the proposed approach can significantly improve the synthesis accuracy on unseen domains in a 3D setting. The code is publicly available at https://github.com/WinstonHuTiger/2D_VAE_UDA_for_3D_sythesis.

1 Introduction

Medical image synthesis is drawing increasing attention in medical imaging, because it could generate missing image data, improving diagnosis and benefits many downstream tasks such as image segmentation [3,5,16]. For example, missing modality is a common issue in multi-modal neuroimaging, e.g., due to motion in the acquisition process [2]. However, existing synthesis frameworks

Q. Hu and H. Li—Equal contributions to this work.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-16446-0_47.

L. Wang et al. (Eds.): MICCAI 2022, LNCS 13436, pp. 495–504, 2022.
https://doi.org/10.1007/978-3-031-16446-0_47

are mostly developed and evaluated on single-domain data (e.g., images from the same scanner), with limited consideration of model robustness when testing on unseen image domains which might be collected from another imaging scanner or acquisition protocols. Hence, domain adaptation is crucial for the real-world deployment in clinical routine. In particular, unsupervised domain adaptation (UDA) is more practical as it does not require additional expensive supervision to fine-tune the pre-trained model.

It should be noted that UDA of classification [1,11,17] and segmentation [4,9,10,12,15] models are well explored in recent years. For image segmentation models, the problem formulation is as follows. Given two different *input domains* (i.e., source and target) with data X and its distribution $P(X)$, $\mathcal{D}_s = \{X_s, P(X_s)\}$, $\mathcal{D}_t = \{X_t, P(X_t)\}$ and a shared *output* space $\mathcal{Y} = \{Y\}$, a predictive model $f(\cdot)$ which approximates $P(Y|X)$ trained on the source domain $\mathcal{D}_s$ is likely to degrade on the target domain $\mathcal{D}_t$ which presents a domain shift. Among existing works, one of the key ideas is to match the *input space* for both domains in the feature space so that the mapping can be invariant to the inputs. It could be achieved by adversarial training [17] or prior matching [14].

As shown in Fig. 1, in both classification and segmentation tasks, the output label spaces in source and target domain are shared. For example, a segmentation model segments the same anatomical structure in both domains. However, in a synthesis model, the *output* spaces from source domain Y_s and target domain Y_t are most likely different, for example, the outputs images Y_s and Y_t could be from different scanners. In the UDA scenario, we only have access to the input of target domain X_s, thus matching the synthetic output $\hat{Y}_t$ to its real distribution is challenging as there is no observations of the outputs. Importantly, aligning X_t and X_s does not guarantee that the output would be close to Y_t but Y_s. Thus, most existing works in classification and segmentation could not be directly applied to synthesis model. Generally, we expect the generated output of the target domain $\hat{Y}_t$ to match a *reasonable* distribution of the target domain. In this work, we present the problem setting, upper bound and propose an efficient approach to perform UDA in a 3D setting.

Why 3D-UDA Is Necessary and Challenging? Previous work focusing on 2D or patch-based adaptation [4,8]. Although these works show promising results, they are limited to 2D or patch domains which is insufficient for many applications such as neuroimaging data which requires domain adaptation in a 3D fashion. The 3D image-to-image synthesis model dealing with full-volume imaging data is heavy-weight compared to patch-based method. However, extending existing work from 2D to 3D is non-trivial. In addition to model complexity, another challenge is that the number of 3D volumetric samples is very limited while 2D slices are more accessible.

Contributions. Our contribution is threefold: (1) We introduce unsupervised domain adaptation for 3D medical image synthesis and present the technical difference with existing setup in image classification and segmentation. (2) We

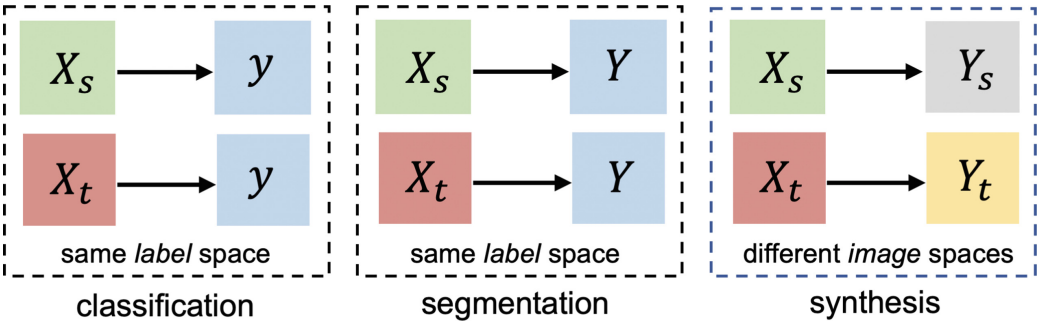


Fig. 1. Summary of the main differences in domain adaptation between image classification, segmentation, and synthesis tasks. The output spaces with the same colors indicate the output spaces have the same distribution. For example, a segmentation model segments the same anatomical structure in both domains.

propose an efficient 2D variational-autoencoder approach to perform UDA in a 3D manner. (3) We present empirical studies on the effects of the amount of data needed for adaptation and the effect of key hyper-parameters. Our results show that the proposed method can significantly improve the synthesis accuracy on unseen domains.

2 Methodology

Problem Definition. The objective is to adapt an volume-to-volume mapping Φ_s: $X_s \rightarrow Y_s$ which is trained on a source domain to a target domain, so that when testing on input data X_t, the output could match the target distribution: Let $\mathcal{S}$ and $\mathcal{T}$ denote the source domain and the target domain, respectively. We observe N samples $S = \{(x_s^k, y_s^k)\}_{k=1}^N$ from $\mathcal{S}$, and M samples $T = \{x_t^j\}_{j=1}^M$ from $\mathcal{T}$. Notably, the samples from the target domain do not contain any output data.

Supervised Domain Adaptation. When there is some target data $\{X_t, P(X_t)\}$ available, one could use them to fine-tune the established mapping M and transfer the knowledge from source to target. When increasing the amount of data for tuning, the upper bound could be setup for *unsupervised* domain adaptation in which only the input data from the target domain can be accessible.

Unsupervised Domain Adaptation. In this setting, X_t is available while Y_t is not accessible. Since the goal of a synthesis model is to generate *reasonable* output. One straightforward approach is to match the 3D prior distributions of $\hat{Y}_t$ and Y_s. Although Y_s and Y_t are expected to be different, they largely share the underlying distribution, e.g., images from different scanners may present varied contrasts but share the same space of anatomical structure. However, directly modeling 3D distribution with limited data is challenging. As an alternative, we explore to model the 3D distribution with a 2D spatial variational autoencoder (s-VAE) which is effective and computationally efficient.

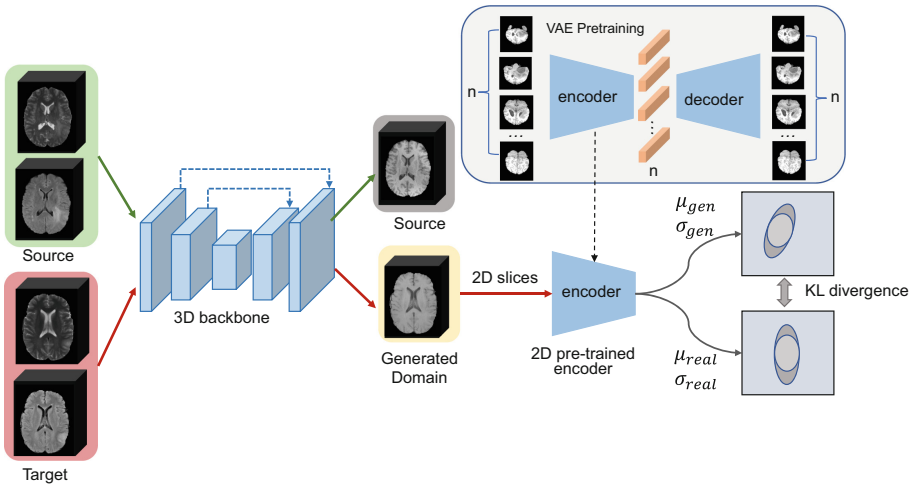


Fig. 2. The main framework of our proposed 3D-UDA method for cross-modality MR image synthesis. In the source domain (green), the 3D backbone is trained supervisedly with aligned image data, translating FLAIR and T2 to T1. A spatial 2D variational autoencoder is first trained in the source domain using T1 modality. The dimension of its latent space is n. Then, in the adaptation stage, we compute the KL-divergence between the prior distribution of the generated volume and the target 3D distribution learned by a 2D variational autoencoder. (Color figure online)

2D s-VAE for Modeling 3D Distribution. To encode 3D distribution in the 2D VAE's latent space, we proposed to train the VAE in a volumetric way, i.e., instead of training the 2D VAE with slices from different brain volumes, we take shuffled 2D slices from *a whole 3D volume* as the input. Thus, the batch size corresponds to the number of slices in the volume. We nickname such a form of VAE as spatial VAE (s-VAE). Doing this ensures that the 2D s-VAE learns the correlation between 2D slices. Since each latent code comes from a specific slice of a whole 3D brain volume, n latent codes with a certain sequence together can express the 3D distribution, while a standard 3D VAEs encode the distribution in their channel dimension. 2D s-VAE can reduce learnable parameters compared to 3D VAEs. The training of s-VAE is formulated as:

$$\mathcal{L}_{VAE} = D_{KL}(N(0, I)||N(\mu_{\hat{Y}_s}, \sigma_{\hat{Y}_s})) + ||Y_s - \hat{Y}_s||_2 \tag{1}$$

Backbone 3D ConvNets for Image Synthesis. One basic component is a 3D ConvNets backbone for image synthesis. With the N paired samples from the source domain, supervised learning was conducted to establish a mapping Φ_s from the input image space X_s to the output space Y_s, optimized with an L1 norm loss: $\mathcal{L}_{syn} = \sum_{i=1}^{N} ||x_s^i - x_t^i||_1$.

3D UDA with 2D s-VAE. Once the 3D distribution of the output in source domain $P(Y_s)$ is learned by 2D s-VAE, we could match it with the posterior

distribution $P(\hat{Y}_t)$ given the generated output $\hat{Y}_t$. Kullback-Leibler (KL) divergence is employed to match $P(\hat{Y}_t)$ and $P(Y_s)$, which can be formulated as

$$\mathcal{L}_{ada} = D_{KL}(P(Y_s)||P(\hat{Y}_t)) = \sum P(Y_s) log(\frac{P(Y_s)}{P(\hat{Y}_t)}) \tag{2}$$

However, just optimizing the KL divergence can be problematic since the mapping from input to output might suffer from catastrophic forgetting. Consequently, we perform supervised training on the source domain while adapting it to target domain with KL divergence. The whole UDA pipeline is shown in Fig. 2.

3 Experiments

Datasets and Preprocessing. We use the multi-center *BraTS* 2019 dataset [13] to perform cross-modality image-to-image synthesis and investigate domain adaptation. Specifically, we generate T1 images from the combination of FLAIR and T2 images. To create source and target domains for *UDA*, we split it into two subsets based on the IDs of medical centers. The first subset contains 146 paired multi-modal MR images from *CBICA* while the second subset consists of 239 paired MR images from *TCIA*. In the supervised image-to-image synthesis training process and domain adaptation process, input volumes loaded as gray scaled image are first cropped from (155, 240, 240) to (144, 192, 192) to save memory cost. Then three data argumentation methods are applied to the cropped volumes, including random rotation, random flip and adjusting contrast and brightness. The contrast level and brightness level are chosen from a uniform distribution $(0.3, 1.5]$.

In our setting, there are two domain adaptation tasks: CBICA $\rightarrow$ TCIA; TCIA $\rightarrow$ CBICA. Specifically, in the first task, the image synthesis model is trained in a supervised manner using CBICA and then the model is adapted to TCIA subset. The second task is the reverse of the adaptation direction.

In the supervised training, all data from the source domain are used; in the unsupervised domain adaptation stage, we utilize 100 input volumes (FLAIR and T2) from the target domain without any output volumes (T1). The rest of volumes from the target domain are then evaluating the performance of all methods. The details of composition of datasets are summarized in Table 1.

Table 1. The composition of datasets for the settings of three scenarios in two domain adaptation tasks. (Adapt. = Adaptation)

Methods	CBICA $\rightarrow$ TCIA			TCIA $\rightarrow$ CBICA		
	Source	Target		Source	Target	
		Adapt. set	Test set		Adapt. set	Test set
Without DA	146	0	139	146	0	46
UDA	146	100	139	239	100	46
Supervised DA	0	[40, 100]	139	0	[40, 100]	46

Configuration of the Training Schedule. We use a 3D pix2pix model [7] as the pipeline to perform cross-modality synthesis. The generator of the pix2pix model has a large $9 \times 9 \times 9$ receptive field and it has only four down-sampling stage to reduce computation complexity. FLAIR and T2 volumes are concatenated into a two-channel input. For the two domains, we train the 2D s-VAE model individually using T1 volumes from each of the source domain. A single volume is re-scaled from (240, 240, 155) dimension to 256 slices with dimension (256, 256). 2D s-VAE is trained for 300 epoch and the KL divergence loss weight is set to be 0.001. For the synthesis model, we first train the model using source dataset in a supervised way. Then, in the UDA process, we train the model for five epochs. In the first iteration of the UDA process, we perform supervised training on the source domain with previous hyper-parameters; in the second iteration, we fine-tune the 3D backbone with the pre-trained 2D s-VAE model. All models are trained on RTX 3090 with Pytorch 1.10.2. Due to page limit, details of the backbone synthesis architecture, 2D VAE and 3D VAE are shown in the Appendix.

Evaluation Protocol. We use structural similarity index (SSIM) [19] and peak signal-to-noise ratio (PSNR) to evaluate the image quality of the synthesized images. We further use a pre-trained nnUnet [6] on the *BraTS* 2020 dataset [13] to segment the generated images from different methods and report Dice scores. All the test images are not involved in the UDA process.

Table 2. Comparison of our methods with the lower bound, upper bound, two baselines, supervised method without domain shift and real images (only on Dice). The p-values for: a) ours vs. lower bound, and b) ours vs. 3D-VAE are all smaller than 0.0001, indicating our method outperforms the two methods significantly. We only report the mean of Dice score due to page limit.

Methods	CBICA →TCIA			TCIA → CBICA		
	SSIM	PSNR	Dice	SSIM	PSNR	Dice
Without DA (lower bound)	0.837 (± 0.030)	19.999 (± 2.150)	0.793	0.837 (± 0.027)	18.976 (± 3.685)	0.871
Without DA+augmentation	0.845 (± 0.028)	21.525 (± 2.204)	0.774	0.833 (± 0.029)	19.101 (± 3.646)	0.874
3D VAE UDA	0.844 (± 0.026)	21.183 (± 2.252)	0.772	0.832 (± 0.029)	19.278 (± 3.611)	0.874
2D s-VAE UDA (Ours)	**0.853** (± 0.024)	**22.217** (± 2.253)	0.773	**0.846** (± 0.024)	**19.591** (± 3.429)	0.874
Supervised DA ($n = 10$)	0.844 (± 0.024)	24.969 (± 2.634)	0.763	0.851 (± 0.014)	22.509 (± 2.062)	0.864
Supervised DA ($n = 40$)	0.869 (± 0.026)	24.933 (± 2.828)	0.790	0.852 (± 0.017)	23.811 (± 2.365)	0.866
Supervised DA ($n = 100$)	0.869 (± 0.026)	24.619 (± 2.953)	0.799	0.865 (± 0.017)	23.877 (± 2.611)	0.870
Upper bound	0.911 (± 0.0263)	25.519 (± 3.630)	0.820	0.896 (± 0.020)	24.656 (± 2.907)	0.867
Real Images	-	-	0.904	-	-	0.936

4 Results

4.1 Comparison of Methods

We first present the lower bound of the synthesis model without DA on the target domain. Then we present our proposed 2D s-VAE method. We also present

the upper bound of a supervised DA. Quantitative results are summarized in Table 2. Moreover, we study the impact of the amount of volumes used in the UDA process and the impact of different batch sizes in 2D VAE reconstruction results, showed in Fig. 4.

Lower-Bound and Upper-Bound. The lower-bound of the UDA is pre-training on source domain and directly testing on the target domain. Notably basic data argumentation like random rotation rotation, random flipping are used to prevent models from over-fitting. As we can observe from row 1 of Table 2, it achieves the worst performance among all the methods. Given available paired data from the target domain, one could tune the pre-trained model (same as transfer learning) to adapt the model to the target domain. One could observe that increasing data from target domain for supervision improves the performance, from the three rows ('Supervised DA') in Table 2. The upper bound is defined by training and evaluating the method directly on the target domain. The last second row in Table 2 shows the results of five-fold cross validation on the target domain.

Heuristic Data Argumentation. Heuristic data argumentation could potentially improve the generalizability [20]. We perform contrast and brightness-related data augmentation considering that one of the most important domain shift is the image contrast between different scanners. We observe that it brings slight improvement in adaptation by comparing row 1 and row 2 in Table 2.

3D VAE *vs.* 2D s-VAE. As another baseline to be compared with our proposed efficient 2D s-VAE approach, we trained 3D VAEs using the volumetric data from the source domains. One could observe that the 3D VAE performs comparably with the heuristic data argumentation approach. This is partly because there are limited data to train the 3D VAE for learning a proper underlying distribution.

Our proposed 2D s-VAE method outperforms both data augmentation and the 3D VAE method on both SSIM and PSNR in two tasks. 3D VAE encoder is more computationally expensive, since the encoder of 3D VAE has 5.17M learnable parameters while 2D s-VAE only has 1.73M ones. Although there is still a visible performance gap between all the UDA methods and the upper bound, our 2D s-VAE method provides an effective and efficient solution when the output modality from the target domain is not accessible.

4.2 Analysis of Parameters

Impact of Batch Size on 2D s-VAE: In 2D s-VAE training process and the UDA process, slices of a whole brain MRI sequence serve as the input. To show the impact of batch size on 2D s-VAE, we explore the batch size of values, 32, 64, 128 and 256. To understand how much the VAE models the 3D distribution and visualize the results from the learned prior, we follow [18] to build a Guassian model of 2D slices, which models the correlation of 2D slices in the latent space

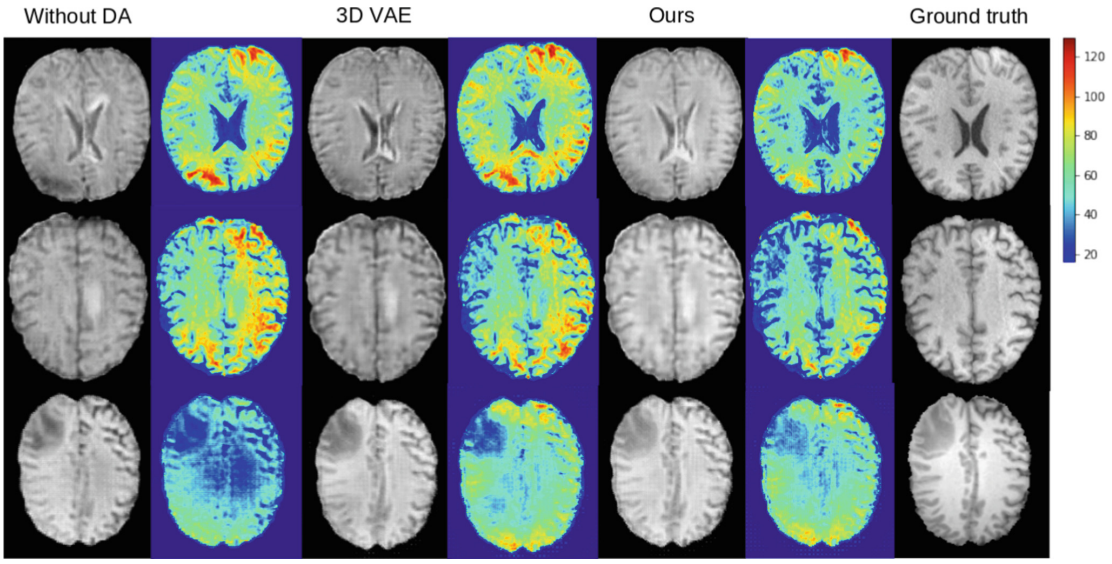


Fig. 3. Results on example slices achieved by different methods: (a) Without domain adaptation (DA), (b) 3D VAE, and (c) our 2D s-VAE. The difference map (in heatmap) is computed by subtracting the ground truth slice and the synthesized slice. We observe that our approach achieves best perceptual quality.

for sampling infinite number of 2D slices. As we show in Fig. 4(b), when batch size is 32, almost nothing can be sampled from the learned distribution; when batch size is 64, only noise can be sampled; when batch size is 128 (at this point half of the brain slices are used in a batch), brain shape can be visualized in the samples; when batch size is 256, brain structures are clear in the samples.

Impact of the Amount of Volumes: As we show in Fig. 4(b), we study the impact of amount of volumes used for the UDA process. We observe that for both CBICA → TCIA and TCIA → CBICA tasks, when the number of volumes is less than 70, the performance increases. However, when the number exceeds 70,

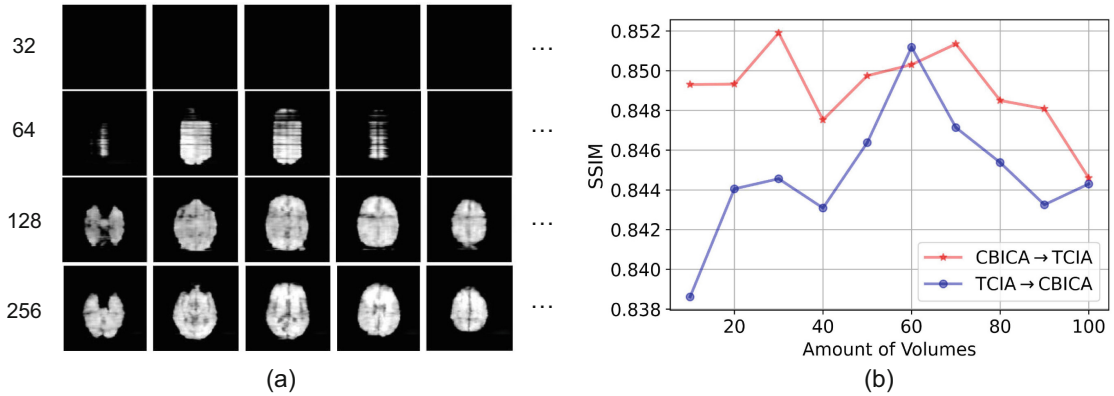


Fig. 4. (a): the reconstruction results influenced by different batch sizes. Larger batch size could better capture the 3D distribution. (b): the effect of the amount of volumes used for UDA from the target domain.

the performance starts to decrease. That is because the first continuing training batch in the UDA process contributes more to the results rather than the second batch. Although the first batch regulates the whole UDA process, it might hurt the performance in some degree.

5 Discussion

In this work, for the first time, we explore domain adaptation for medical image-to-image synthesis models. We first explain the difference between the domain adaptation of synthesis, classification and segmentation models. Then we introduce our efficient unsupervised domain adaptation method using 2D s-VAE when the target domain is not accessible. Finally, we show the effectiveness of our 2D s-VAE method and study some factors that influence the performance of our framework. In our approach, we translate the whole volume from one domain to another instead of using a patch-based method. Although whole-volume approaches are able to capture the full spatial information, it suffers from limited training data issues. As we have shown in Fig. 3, even after domain adaptation, we observed that the domain gap is challenging to overcome. Recent disentangled learning that could separate domain-specific and shared features effectively might improve the current results. Contrastive learning could be explored to better capture the representation of the source or target domains more effectively. Given the above limitations, we still wish our approach provides a new perspective for robust medical image synthesis for future research.

Acknowledgement. This work is supported in part by National Key Research and Development Program of China (No.: 2021YFF1200800) and Shenzhen Science, Technology and Innovation Commission Basic Research Project (No. JCYJ20180507181527806). H. L. was supported by Forschungskredit (No. FK-21-125) from UZH.

References

1. Ahn, E., Kumar, A., Fulham, M., Feng, D., Kim, J.: Unsupervised domain adaptation to classify medical images using zero-bias convolutional auto-encoders and context-based feature augmentation. IEEE Trans. Med. Imaging **39**(7), 2385–2394 (2020)
2. Conte, G.M., et al.: Generative adversarial networks to synthesize missing T1 and FLAIR MRI sequences for use in a multisequence brain tumor segmentation model. Radiology **300**(1), E319 (2021)
3. Conte, G.M., et al.: Generative adversarial networks to synthesize missing T1 and flair MRI sequences for use in a multisequence brain tumor segmentation model. Radiology **299**(2), 313–323 (2021)
4. Dou, Q., Ouyang, C., Chen, C., Chen, H., Heng, P.A.: Unsupervised cross-modality domain adaptation of convnets for biomedical image segmentations with adversarial loss. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence, pp. 691–697 (2018)

5. Finck, T., et al.: Deep-learning generated synthetic double inversion recovery images improve multiple sclerosis lesion detection. Investig. Radiol. **55**(5), 318–323 (2020)
6. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: NNU-net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021)
7. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5967–5976 (2017)
8. Kamnitsas, K., et al.: Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In: Niethammer, M., et al. (eds.) IPMI 2017. LNCS, vol. 10265, pp. 597–609. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59050-9_47
9. Kamnitsas, K., et al.: Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In: Niethammer, M., et al. (eds.) IPMI 2017. LNCS, vol. 10265, pp. 597–609. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59050-9_47
10. Karani, N., Erdil, E., Chaitanya, K., Konukoglu, E.: Test-time adaptable neural networks for robust medical image segmentation. Med. Image Anal. **68**, 101907 (2021)
11. Lenga, M., Schulz, H., Saalbach, A.: Continual learning for domain adaptation in chest x-ray classification. In: Medical Imaging with Deep Learning, pp. 413–423. PMLR (2020)
12. Liu, L., Zhang, Z., Li, S., Ma, K., Zheng, Y.: S-CUDA: self-cleansing unsupervised domain adaptation for medical image segmentation. Med. Image Anal. **74**, 102214 (2021)
13. Menze, B.H., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE Trans. Med. Imaging **34**(10), 1993–2024 (2014)
14. Ouyang, C., Kamnitsas, K., Biffi, C., Duan, J., Rueckert, D.: Data efficient unsupervised domain adaptation for cross-modality image segmentation. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11765, pp. 669–677. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32245-8_74
15. Perone, C.S., Ballester, P., Barros, R.C., Cohen-Adad, J.: Unsupervised domain adaptation for medical imaging segmentation with self-ensembling. NeuroImage **194**, 1–11 (2019)
16. Thomas, M.F., et al.: Improving automated glioma segmentation in routine clinical use through artificial intelligence-based replacement of missing sequences with synthetic magnetic resonance imaging scans. Investig. Radiol. **57**(3), 187–193 (2022)
17. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7167–7176 (2017)
18. Volokitin, A., et al.: Modelling the distribution of 3D brain MRI using a 2D slice VAE. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12267, pp. 657–666. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59728-3_64
19. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
20. Zhang, L., et al.: Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. IEEE Trans. Med. Imaging **39**(7), 2531–2540 (2020)

7 Imbalance-aware self-supervised radiomics

This chapter has been published as one **peer-reviewed conference publication**:

[1] **H. Li**, F.-F. Xue, K. Chaitanya, S. Luo, I. Ezhov, B. Wiestler, J. Zhang, and B. Menze. "Imbalance-aware self-supervised learning for 3d radiomic representations". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (**MICCAI**)*. Springer. 2021, pp. 36–46

Synopsis: This work develops a novel data-driven radiomics and address the imbalance issue in self-supervised learning. Technically, a self-supervised representation learning framework is established to learn high-level features of 3D volumes as a complement to existing radiomics features. Specifically, a 3D Siamese network is developed to learn image representations in a self-supervised fashion. More importantly, implicit data imbalance is addressed by exploiting two unsupervised strategies: a) sample re-weighting, and b) balancing the composition of training batches. When combining the learned self-supervised feature with traditional radiomics, significant improvement is shown in brain tumor classification and lung cancer staging tasks covering MRI and CT imaging modalities.

Contributions of thesis author: algorithm design and implementation, computational experiments and composition of manuscript.

Imbalance-Aware Self-supervised Learning for 3D Radiomic Representations

Hongwei Li[1,2], Fei-Fei Xue[4], Krishna Chaitanya[3], Shengda Luo[5], Ivan Ezhov[1], Benedikt Wiestler[6], Jianguo Zhang[4(✉)], and Bjoern Menze[1,2]

[1] Department of Computer Science, Technical University of Munich, Munich, Germany
hongwei.li@tum.de
[2] Department of Quantitative Biomedicine, University of Zurich, Zürich, Switzerland
[3] ETH Zurich, Zürich, Switzerland
[4] Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, China
zhangjg@sustech.edu.cn
[5] Faculty of Information Technology, Macau University of Science and Technology, Macao, China
[6] Klinikum rechts der Isar, Technical University of Munich, Munich, Germany

Abstract. Radiomics can quantify the properties of regions of interest in medical image data. Classically, they account for pre-defined statistics of shape, texture, and other low-level image features. Alternatively, deep learning-based representations are derived from supervised learning but require expensive annotations and often suffer from overfitting and data imbalance issues. In this work, we address the challenge of learning the representation of a 3D medical image for an effective quantification under data imbalance. We propose a *self-supervised* representation learning framework to learn high-level features of 3D volumes as a complement to existing radiomics features. Specifically, we demonstrate how to learn image representations in a self-supervised fashion using a 3D Siamese network. More importantly, we deal with data imbalance by exploiting two unsupervised strategies: a) sample re-weighting, and b) balancing the composition of training batches. When combining the learned self-supervised feature with traditional radiomics, we show significant improvement in brain tumor classification and lung cancer staging tasks covering MRI and CT imaging modalities. Codes are available in https://github.com/hongweilibran/imbalanced-SSL.

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-87196-3_4) contains supplementary material, which is available to authorized users.

M. de Bruijne et al. (Eds.): MICCAI 2021, LNCS 12902, pp. 36–46, 2021.
https://doi.org/10.1007/978-3-030-87196-3_4

1 Introduction

Great advances have been achieved in supervised deep learning, reaching expert-level performance on some considerably challenging applications [11]. However, supervised methods for image classification commonly require relatively large-scale datasets with ground-truth labels which is time- and resource-consuming in the medical field. Radiomics is a translational field aiming to extract objective and quantitative information from clinical imaging data. While traditional radiomics methods, that rely on statistics of shape, texture and others [1], are proven to be generalizable in various tasks and domains, their discriminativeness is often not guaranteed since they are low-level features which are not specifically optimized on target datasets.

Self-supervised learning for performing *pre-text tasks* have been explored in medical imaging [24,25], that serve as a proxy task to pre-train the deep neural networks. They learn representations commonly in a supervised manner on proxy tasks. Such methods depend on heuristics to design pre-text tasks which could limit the discriminativeness of the learnt representations. In this work, we investigate self-supervised *representation learning* which aims to **directly** learn the representation of the data without a proxy task.

Recent contrastive learning-based methods [6,15,20] learn informative representations *without* human supervision. However, they often rely on large batches to train and most of them work for 2D images. To this end, due to the high dimensionality and limited number of training samples in medical field, applying contrastive learning-based methods may not be practically feasible in 3D datasets. Specially, in this study, we identify two main differences required to adapt self-supervised representation learning for radiomics compared to natural image domain: i) Medical datasets are often multi-modal and three dimensional. Thus, learning representation methods in 3D medical imaging would be computationally expensive. ii) heterogeneous medical datasets are *inherently imbalanced*, e.g. distribution disparity of disease phenotypes. Existing methods are built upon approximately balanced datasets (e.g. CIFAR [18] and ImageNet [10]) and do not assume the existence of data imbalance. Thus, how to handle data imbalance problem is yet less explored in the context of self-supervised learning.

Related Work. Radiomic features have drawn considerable attention due to its predictive power for treatment outcomes and cancer genetics in personalized medicine [12,23]. Traditional radiomics include shape features, first-, second-, and higher- order statistics features.

Self-supervised representation learning [3,6,7,13,15,22] have shown steady improvements with impressive results on multiple natural image tasks, mostly based on contrastive learning [14]. Contrastive learning aims to attract positive (or *similar*) sample pairs and rebuff negative (or *disimilar*) sample pairs. Positive sample pairs can be obtained by generating two augmented views of one sample, and the remaining samples in the batch can be used to construct the negative samples/pairs for a given positive pair. In practice, contrastive learning

methods benefit from a large number of negative samples. In medical imaging, there are some existing work related to contrastive learning [2,17,21]. Chaitanya *et al.*'s work [5] is the most relevant to our study, which proposed a representation learning framework for image segmentation by exploring local and global similarities and dissimilarities. Though these methods are effective in learning representations, they require a large batch size and/or negative pairs, which make them difficult to apply to *3D* medical data. Chen *et al.* [8] demonstrates that a Siamese network can avoid the above issues on a 2D network. The Siamese network, which contains two encoders with shared weights, compares two similar representations of two augmented samples from one sample. Importantly, it neither uses negative pairs nor a large batch size. Considering these benefits, we borrow the Siamese structure and extend it to 3D *imbalanced* medical datasets.

Contributions. Our contribution is threefold: (1) We develop a 3D Siamese network to learn self-supervised representation which is high-level and discriminative. (2) For the first time, we explore how to tackle the data imbalance problem in self-supervised learning without using labels and propose two effective unsupervised strategies. (3) We demonstrate that self-supervised representations can complement the existing radiomics and the combination of them outperforms supervised learning in two applications.

2 Methodology

The problem of interest is how to learn high-level, discriminative representations on 3D imbalanced medical image datasets in a self-supervised manner. The schematic view of the framework is illustrated in Fig. 1. First, a pre-trained 3D encoder network, denoted as E_a, takes a batch of original images X with batch size N as input and outputs N representation vectors. The details of the 3D encoder is shown in Table 4 of the Supplementary. The features are fed into the *RE/SE* module to estimate their individual weights or to resample the batch.

Then each image x in the batch X is randomly augmented into two images (or called an image pair). They are processed and compared by a 3D Siamese network, nicknamed *3DSiam*, which enjoys relatively low memory without relying on large training batch of 3D data. The proposed *3DSiam* extends original 2D Siamese network [8] from processing 2D images to 3D volumes while inherits its advantages. Since medical datasets are inherently imbalanced, by intuition sole *3DSiam* would probably suffer from imbalanced data distribution. In the following, we first introduce *RE/SE* module to mitigate this issue.

RE/SE Module to Handle Imbalance. Since there is no prior knowledge on the data distribution available, the way to handle imbalance must be *unsupervised*. The vectors mentioned above are fed into a *RE/SE* module before training the *3DSiam* network. The k-means algorithm is used first to cluster the representation vectors into k centers. We then proposed two simple yet effective strategies: a) sample re-weighting (RE), and b) sample selection (SE):

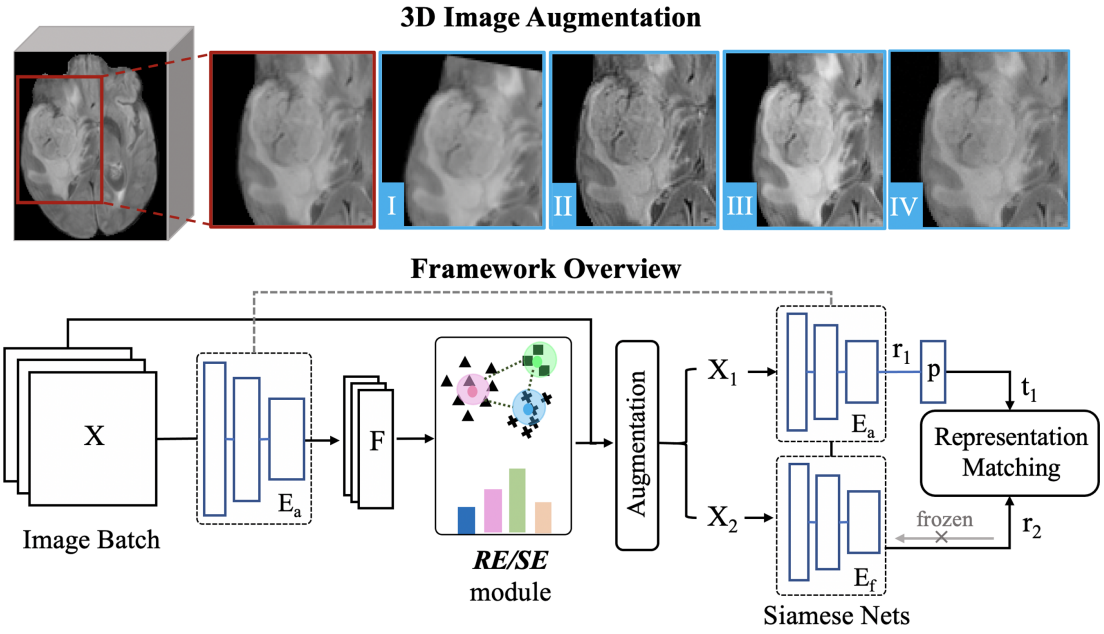


Fig. 1. Our proposed framework learns invariance from extensive 3D image augmentation within four categories: I) affine transform, II) appearance enhancement, III) contrast change, and IV) adding random noise. First, an image batch X is fed into an initialized 3D encoder to obtain its representation F. The RE/SE module first estimates its distribution by k-means based clustering and uses two strategies including sample re-weighting (RE) or sample selection (SE) to alleviate data imbalance issue. Each image is randomly augmented into two positive samples $\{X_1, X_2\}$ which are then used to train a 3D Siamese network by comparing their representations from two encoders $\{E_a, E_f\}$ with shared weights. p is a two-layer perceptron to transform the feature.

a) Sample re-weighting (RE). Denote a batch with N samples as $X = \{x_i | i = 1, 2, ..., N\}$. Given k clusters, denote the distribution of k clusters of features as $F = \{f_j | j = 1, 2, ..., k\}$ over N samples. f_j denotes the frequency of cluster j. Then we assign different weights to the samples in each cluster j. For each sample x_i, representation vector of which belongs to cluster j, we assign it a weight of N/f_j to penalize the imbalanced distribution during the batch training. In practice, we further normalize it by re-scaling to guarantee the minimum weight is 1 for each input batch.

b) Sample selection (SE). Denoting the clusters' centroids as $C = \{c_1, c_2, ...c_k\}$, we find the maximum Euclidean distance $max_{i,j \in [1,k], i \neq j} d(c_i, c_j)$ among all pairs of centroids. k is a hyper-parameter here. We hypothesize that the clusters with maximum centroid distance are representation vectors from different groups. To select m samples from the original N samples to form a new batch, denoted by $B_c = \{x_1, x_2, ..., x_m\}$, we sample $\frac{m}{2}$ nearest sample points centered on each of the selected maximum-distance centroids. m is set to be smaller than $\frac{N}{k}$ for low computation complexity and efficient sampling. The selected new batch is then used to train our *3DSiam* network. A motivation behind the selection strategy is outlined in Supplementary.

3D Siamese Network. The *3DSiam* takes as input two randomly augmented views x_1 and x_2 from a sample x. The two views are processed by two 3D encoder networks with *shared* weights. One of the encoder has frozen weights when training (denoted as E_f) and the other one is with active weights (denoted as E_a). Before training, E_f is always updated to the weights of E_a. E_a is followed by a two-layer perceptron called *predictor* to transform the features. The final objective is to optimize a matching score between the two *similar* representations $t_1 \triangleq p\left(E_a(x_1)\right)$ and $r_2 \triangleq E_f\left(x_2\right)$. *3DSiam* minimizes their negative cosine similarity, which is formulated as:

$$S\left(t_1, r_2\right) = -\frac{t_1}{\|t_1\|_2} \cdot \frac{r_2}{\|r_2\|_2}, \tag{1}$$

where $\|\cdot\|_2$ is L_2-norm. Following [6], we define a symmetrized loss to train the two encoders, formulated as:

$$\mathcal{L} = \frac{1}{2} S\left(t_1, r_2\right) + \frac{1}{2} S\left(t_2, r_1\right), \tag{2}$$

where $t_2 \triangleq p\left(E_f(x_2)\right)$, $r_1 \triangleq E_f(x_1)$. This loss is defined and computed for each sample with re-weighting in the batch X or the new batch B_c with equal weights. Notably the encoder E_f on x_2 receives no gradient from r_2 in the first term of Eq. (2), but it receives gradients from t_2 in the second term (and vice versa for x_1). This training strategy avoids collapsing solutions, i.e., t_1 and r_2 outputs a constant over the training process. When training is finished, r_2 is used as the final representation.

3 Experiments

Datasets and Preprocessing. The evaluation of our approach is performed on two public datasets: 1) a multi-center MRI dataset (*BraTS*) [4,19] including 326 patients with brain tumor. The MRI modalities include FLAIR, T1, T2 and T1-c with a uniform voxel size $1 \times 1 \times 1\,\mathrm{mm}^3$. Only FLAIR is used in our experiment for simplicity of comparisons. 2) a lung CT dataset with 420 non-small cell lung cancer patients (*NSCLC-radiomics*) [1,9][1]. The effectiveness of the learnt representations is evaluated on two classification tasks (also called 'down-stream task'): a) discriminating high grade (H-grade) and low grade tumor (L-grade), and b) predicting lung cancer stages (i.e. I, II or III). The *BraTS* dataset is imbalanced in two aspects: a) the distribution of ground truth labels (H-grade *vs.* L-grade); b) the distribution of available scans among different medical centers. For *NSCLC-radiomics*, the distribution of ground truth labels are imbalanced as well, with ratio of 2:1:6 for stage I, II and III respectively. For *BraTS*, we make use of the segmentation mask to get the centroid and generate a 3D bounding box of $96 \times 96 \times 96$ to localize the tumour. If the bounding box exceeds the original volume, the out-of-box region was padded with background

[1] Two patients were excluded as the ground truth labels are not available.

intensity. For *NSCLC-radiomics*, we get the lung mask by a recent public lung segmentation tool [16], and then generate a $224 \times 224 \times 224$ bounding box to localize the lung. The lung volume was then resized to $112 \times 112 \times 112$ due to memory constraint. The intensity range of all image volumes was rescaled to [0, 255] to guarantee the success of intensity-based image transformations.

Configuration of the Training Schedule. We build a 3D convolutional neural network with two bottleneck blocks as the encoder for all experiments (details in Supplementary). In the beginning, we pre-train *3DSiam* for one epoch with a batch size of 6. Then we use it to extract features for the *RE/SE* module. After first epoch, the encoder from the *last iteration* is employed for *dynamic* feature extraction. For down-stream task evaluations, we use the last-layer feature of the encoder. For 3D data augmentation, we apply four categories shown in Fig. 1, including random rotations in $[-20, 20]$ degrees, random scale between $[0.7, 1.3]$, and random shift between $[-0.05, 0.05]$, Gamma contrast adjustment between [0.7, 1.5], image sharpening, and Gaussian blurring, considering the special trait of medical images. For optimization, we use Adam with 10^{-2} learning rate and 10^{-4} weight decay. Each experiment is conducted using one Nvidia RTX 6000 GPU with 24 GB memory. The number of cluster k is set to 3 in all experiments. Its effect is analyzed in the last section.

Computation Complexity. For *3DSiam* without *RE/SE* module, the training takes only around four hours for 50 epochs for the results reported for brain tumor classification task. We do not observe significant improvement when increasing the number of epochs after 50. We train *3DSiam* with *RE/SE* module for around 2000 iterations (not epochs) to guarantee that similar number of training images for the models of comparison are involved. In *RE/SE* module, the main computation cost is from k-means algorithm. We have observed that the overall computation time has increased by 20% (with i5-5800K CPU). It is worth noting that *RE/SE* module is not required during the inference stage, thus there is no increase of the computational cost in testing.

Feature Extraction and Aggregation. For each volume, we extract a set of 107 traditional radiomics features[2] including first- and second-order statistics, shape-based features and gray level co-occurrence matrix, denoted as f_{trad}. For the self-supervised learning one, we extract 256 features from the last fully connected layer of the encoder E_a, denoted as f_{SSL}. To directly evaluate the effectiveness of SSL-based features, we concatenate them to a new feature vector $f = [f_{trad}, f_{SSL}]$. Note that f_{trad} and f_{SSL} are always from the same subjects.

Evaluation Protocol, Classifier and Metrics. For evaluation, we follow the common protocol to evaluate the quality of the pre-trained representations by training a *supervised* linear support vector machine (SVM) classifier on the

[2] https://github.com/Radiomics/pyradiomics.

training set, and then evaluating it on the test set. For binary classification task (BraTS), we use the sensitivity and specificity as the evaluation metrics. For multi-class classification task (lung cancer staging), we report the overall accuracy and minor-class (i.e. stage II) accuracy considering all testing samples. We use *stratified* five-fold cross validation to reduce selection bias and validate each model. In each fold, we randomly sample 80% subjects from each class as the training set, and the remaining 20% for each class as the test set. Within each fold, we employ 20% of training data to optimize the hyper-parameters.

4 Results

Quantitative Comparison. We evaluate the effectiveness of the proposed self-supervised radiomics features on two classification tasks: a) discrimination of low grade and high grade of brain tumor and b) staging of lung cancer.

Table 1. Comparison of the performances of different kinds of features in two downstream tasks using stratified cross-validation. We further reduce 50% training data in each fold of validation to show the effectiveness against supervised learning. Our method outperforms supervised learning in both scenarios.

Methods	*BraTS*		*Lung cancer staging*	
	Sensitivity/Specificity		Overall/Minor-class accuracy	
	Full labels	*50% labels*	*Full labels*	*50% labels*
Trad. radiomics	0.888/0.697	**0.848**/0.697	0.490/0.375	0.481/0.325
Rubik's cube [25]	0.744/0.526	0.680/0.486	0.459/0.325	0.433/0.275
3DSiam	0.844/0.407	0.808/0.526	0.459/0.300	0.445/0.300
3DSiam+SE	0.848/0.513	0.824/0.566	0.471/0.350	0.443/0.325
3DSiam+RE	0.868/0.486	0.828/0.605	0.459/0.375	0.445/0.325
Trad.+3DSiam	0.904/0.645	0.804/0.566	0.495/0.350	0.486/0.350
Trad.+3DSiam+SE	0.916/**0.711**	**0.848/0.763**	**0.538**/0.375	**0.519**/0.350
Trad.+3DSiam+RE	**0.920**/**0.711**	0.804/0.763	0.524/**0.425**	0.502/**0.40**
Supervised learning	0.888/0.711	0.804/0.566	0.526/0.375	0.467/0.325

Effectiveness of RE/SE *Module.* From the first row of Table 1, one can observe that traditional radiomics itself brings powerful features to quantify tumor characteristics. On BraTS dataset, the comparison between traditional radiomics and vanilla self-supervised radiomics (*3DSiam*) confirms our hypothesis that features learned by vanilla self-supervised method behave poorly, especially on the minor class (poor specificity). However, self-supervised radiomics with *RE* or *SE* module surpasses *3DSiam* in specificity by a large margin. The aggregation of the vanilla self-supervised representation and traditional radiomics does not show significant improvement. More importantly, with *RE/SE* module added, the specificity increased by 6.6%, from 64.5% to 71.1%, which indicates a large

boost in predicting the minor class (i.e. L-grade). Both comparisons of rows 4, 5, 6 and rows 7, 8, 9 demonstrate the success of our RE/SE module in tacking class imbalance, i.e., promoting the recognition of the minor class, while preserving the accuracy of the major class.

Comparison with State-of-the-Art. Our method (*Trad.+3DSiam+SE* in Table 2) outperforms the supervised one in two scenarios in two classification tasks, the result of which is achieved by using the same encoder backbone with a weighted cross-entropy loss. When it is trained with 50% less labels, the performance of supervised model decrease drastically. On lung cancer staging with three classes, although the overall accuracy of self-supervised radiomics is lower than the traditional one, with the *RE/SE* module, the combination of two kinds of radiomics achieves the topmost overall accuracy. This demonstrates the proposed self-supervised radiomics is complementary to existing radiomics. In the second row, we show the result of one self-supervised learning method trained by playing Rubik cubes [25] to learn contextual information with a same encoder. We observe that the representation learned in proxy task is less discriminative than the one directly from representation learning.

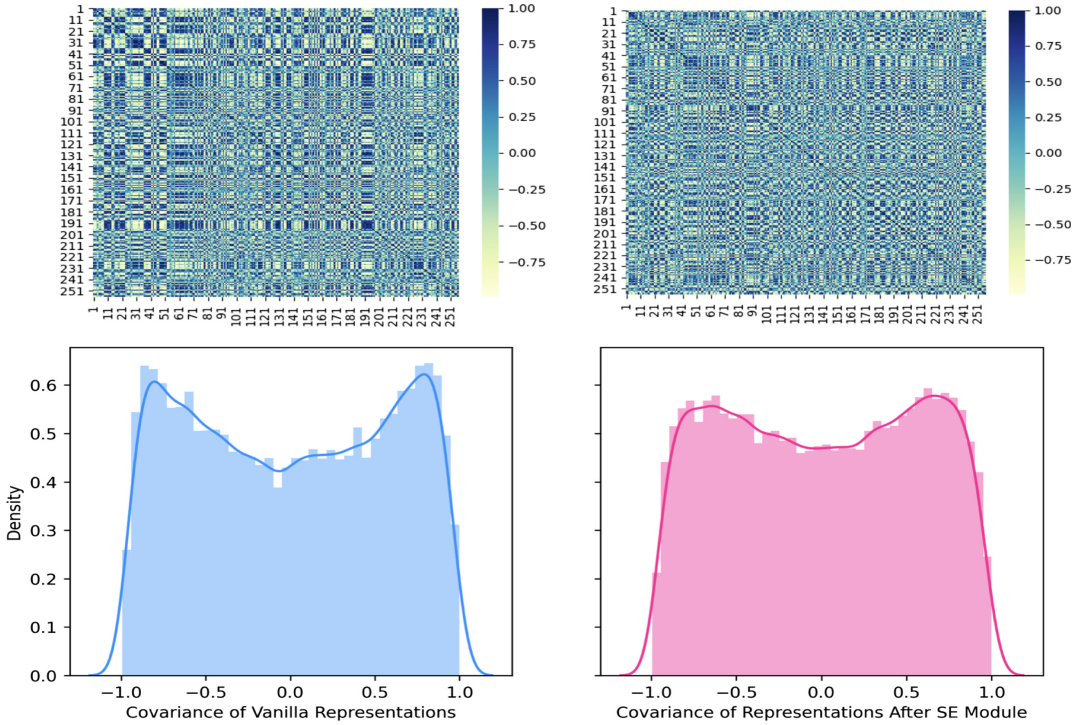


Fig. 2. Covariance analysis of the representations before and after the SE module. Across all 326 tumor patients, each feature was correlated with other ones, thereby generating the correlation coefficients. The density map show that the vanilla representation before SE module are more correlated (redundant) than the one after.

Analysis of Representations and Hyperparameters

Feature Covariance. For a better understanding of the role of the proposed module in relieving data imbalance problem, we further analyze the feature covariance to understand the role of the *SE module*. Consider two paired variables (x_i, x_j) in the representation R. Given n samples $\{(x_{i1}, x_{j1}), (x_{i2}, x_{j2}), ..., (x_{in}, x_{jn})\}$, Pearson's correlation coefficient $r_{x_i x_j}$ is defined as: $r_{x_i x_j} = \frac{cov(x_i, x_j)}{\sigma_{x_i} \sigma_{x_j}}$, where *cov* is the covariance and σ is the standard deviation. We found that the features after *SE* module become more compact as shown in Fig. 2 and more discriminative compared to the features without *SE* module.

Effect of the Number of Clusters k. The hyper-parameter k in the *SE* module is the number of clusters, which plays a vital role in constructing new batch. To evaluate its effect, we use different k to train *3DSiam* and evaluate it through classification task. To fairly compare different values of k, we keep the size m of the new batch B_c fixed to 6 which is also the batch size when $k = 0$ (without *SE* module). The initial batch size N is set to $\mathrm{k} \times \mathrm{q}$ where q is empirically set to 10 in the comparison. The AUC achieves the highest when $k = 3$. With $k = 5$, the AUC drops. This is probably because when k becomes large, the sampling may be biased when only considering a pair of clustering centers. For details, please refer to the curves of AUC over the number of clusters in Table 3 in Supplementary.

5 Conclusion

In this work, we proposed a 3D *self-supervised* representation framework for medical image analysis. It allows us to learn effective 3D representations in a self-supervised manner while considering the imbalanced nature of medical datasets. We have demonstrated that data-driven self-supervised representation could enhance the predictive power of radiomics learned from *large-scale* datasets without annotations and could serve as an effective compliment to the existing radiomics features for medical image analysis. Dealing with imbalance is an important topic and we will explore other strategies in the future.

Acknowledgement. This work was supported by Helmut Horten Foundation. I. E. was supported by the TRABIT network under the EU Marie Sklodowska-Curie program (Grant ID: 765148). S. L. was supported by the Faculty Research Grant (NO. FRG-18-020-FI) at Macau University of Science and Technology. B. W. and B. M. were supported through the DFG, SFB-824, subproject B12. K. C. was supported by Clinical Research Priority Program (CRPP) Grant on Artificial Intelligence in Oncological Imaging Network, University of Zurich.

References

1. Aerts, H.J., et al.: Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat. Commun. **5**(1), 1–9 (2014)
2. Azizi, S., et al.: Big self-supervised models advance medical image classification. arXiv preprint arXiv:2101.05224 (2021)
3. Bachman, P., Hjelm, R.D., Buchwalter, W.: Learning representations by maximizing mutual information across views. In: Advances in Neural Information Processing Systems, pp. 15535–15545 (2019)
4. Bakas, S., et al.: Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. arXiv preprint arXiv:1811.02629 (2018)
5. Chaitanya, K., Erdil, E., Karani, N., Konukoglu, E.: Contrastive learning of global and local features for medical image segmentation with limited annotations. arXiv preprint arXiv:2006.10511 (2020)
6. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. arXiv preprint arXiv:2002.05709 (2020)
7. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020)
8. Chen, X., He, K.: Exploring simple siamese representation learning. arXiv preprint arXiv:2011.10566 (2020)
9. Clark, K., et al.: The cancer imaging archive (TCIA): maintaining and operating a public information repository. J. Digit. Imaging **26**(6), 1045–1057 (2013). https://doi.org/10.1007/s10278-013-9622-7
10. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: Computer Vision and Pattern Recognition, pp. 248–255 (2009)
11. Esteva, A., et al.: Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**(7639), 115–118 (2017)
12. Gillies, R.J., Kinahan, P.E., Hricak, H.: Radiomics: images are more than pictures, they are data. Radiology **278**(2), 563–577 (2016)
13. Grill, J.B., et al.: Bootstrap your own latent-a new approach to self-supervised learning. In: Advances in Neural Information Processing Systems, vol. 33 (2020)
14. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: Computer Vision and Pattern Recognition, pp. 1735–1742 (2006)
15. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Computer Vision and Pattern Recognition, pp. 9729–9738 (2020)
16. Hofmanninger, J., Prayer, F., Pan, J., Röhrich, S., Prosch, H., Langs, G.: Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. Eur. Radiol. Exp. **4**(1), 1–13 (2020). https://doi.org/10.1186/s41747-020-00173-2
17. Kiyasseh, D., Zhu, T., Clifton, D.A.: CLOCS: contrastive learning of cardiac signals. arXiv preprint arXiv:2005.13249 (2020)
18. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
19. Menze, B.H., et al.: The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans. Med. Imaging **34**(10), 1993–2024 (2014)
20. van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)

21. Vu, Y.N.T., Wang, R., Balachandar, N., Liu, C., Ng, A.Y., Rajpurkar, P.: MedAug: contrastive learning leveraging patient metadata improves representations for chest X-ray interpretation. arXiv preprint arXiv:2102.10663 (2021)
22. Ye, M., Zhang, X., Yuen, P.C., Chang, S.F.: Unsupervised embedding learning via invariant and spreading instance feature. In: Computer Vision and Pattern Recognition, pp. 6210–6219 (2019)
23. Yip, S.S., Aerts, H.J.: Applications and limitations of radiomics. Phys. Med. Biol. **61**(13), R150 (2016)
24. Zhou, Z., et al.: Models genesis: generic autodidactic models for 3D medical image analysis. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11767, pp. 384–393. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32251-9_42
25. Zhuang, X., Li, Y., Hu, Y., Ma, K., Yang, Y., Zheng, Y.: Self-supervised feature learning for 3D medical images by playing a Rubik's cube. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11767, pp. 420–428. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32251-9_46

8 Concluding remarks

The increasingly large-scale neuroimaging datasets enable data-driven modeling of domain knowledge. This modeling opens new avenues to improve the diagnosis of neurological diseases by mining non-invasive and reliable biomarkers from advanced MR imaging tools. Although deep learning has enabled substantial leaps in performance in many neuroimage analysis tasks, its deployment in clinics and laboratories remains challenging. A combination of labeling scarcity, missing data, and domain shift between image datasets creates a high adoption barrier and causes underwhelming performance in real-world settings.

This dissertation contributes and validates novel deep learning methods in three neuroimage analysis tasks: (a) image segmentation, (b) image synthesis, and (c) radiomics analysis. These approaches effectively and efficiently process neuroimage data, acknowledge the imperfection of datasets, and compel models to adapt to unseen domains. Due to the publication-based nature of this thesis, Chapter 3 to 7 are self-contained and in their original form. This final chapter, therefore, provides a summary and a more general discussion of the work, including directions for future research.

8.1 Conclusion

In Chapter 3, we developed effective deep-learning and ensemble model-based methods for brain structure segmentation. Specifically, we established a state-of-the-art algorithm for white matter hyperintensities segmentation in FLAIR and T1 scans, which won a grand segmentation challenge at MICCAI 2017. The ensemble of individual single deep neural nets significantly boosts segmentation performance on multi-center MRI scans, especially in improving the segmentation of small lesions. We further developed the methodology for claustrum segmentation in T1-weighted scans by aggregating multi-view information, thus improving segmentation accuracy. Both segmentation methods reach expert-level performance, are evaluated cross-center, and have potential in real-world clinical practice.

In Chapter 4, we explored efficient deep learning to reduce manual annotation efforts for image segmentation tasks. Specifically, we transferred knowledge from *adult* claustrum segmentation in *T1-w* MR scans to *neonate* claustrum segmentation in *T2-w* ones. We demonstrated the effectiveness of the transfer learning technique compared to non-transfer learning approaches. We found that the model could achieve satisfactory segmentation with only 12 annotated scans. Finally, we verified the model's applicability on 528 scans from the DHCP project and revealed reliable segmentations in 97.4

For image synthesis, in Chapter 5, we developed *DiamondGAN*, a unified synthesis framework for multi-contrast MR image synthesis based on generative adversarial networks, and validated it in a clinical setting. Firstly, we developed a unified cross-modality synthesis approach based on generative adversarial networks that can perform arbitrary missing contrast synthesis. Secondly, we applied the developed method to generate synthetic double inversion recovery images and compared their diagnostic performance to conventional sequences in patients with multiple sclerosis (MS). We observed that the generated DIR images improve lesion depiction compared to standard FLAIR and T1 modalities, indicating that we augment the image information **from almost-invisible to visible**.

In Chapter 6, we developed a lesion-specific, uncertainty-aware, and domain-adaptive neuroimage synthesis framework. Firstly, we proposed a new loss function to enhance the image quality

of lesion regions and provide uncertainty estimation. We validated it in a multi-center setting and demonstrated its effectiveness in MS lesion detection compared to the FLAIR modality. Secondly, we explored unsupervised domain adaptation for generative adversarial networks. Importantly, this work highlights the technical differences among the adaptation of image classification, image segmentation, and image synthesis tasks. We developed an efficient domain adaptation approach for 3D image synthesis based on a 2D variational auto-encoder.

Lastly, in Chapter 7, we developed a novel data-driven radiomics approach and addressed the imbalance issue in self-supervised learning. Technically, we established a self-supervised representation learning framework to learn high-level features of 3D volumes to complement existing radiomics features. Specifically, we developed a 3D Siamese network to learn image representations in a self-supervised fashion. More importantly, we addressed an implicit data imbalance issue by exploiting two unsupervised strategies: a) sample re-weighting and b) balancing the composition of training batches. When combining the learned self-supervised feature with traditional radiomics, we observed significant improvement in brain tumor classification and lung cancer staging tasks covering MRI and CT imaging modalities.

8.2 Outlook

In the following sections, we discuss some relevant research topics that could extend the scope of this dissertation.

8.2.1 Segmentation: adaptation, generalization, and deployment

As mentioned in Section 1.2.2, the key difference between domain adaptation and domain generalization lies in whether the model has access to data from the target domain. When data from the target domain is available (either labeled or unlabeled), it becomes possible to learn and obtain characteristics of the target domain to close the domain gap. Supervised domain adaptation (e.g., fine-tuning with labeled data), often akin to transfer learning, has successfully reduced the domain gap with a small set of labeled data, as shown in this dissertation [170] and other works [176]. Unsupervised domain adaptation can be challenging and is not always successful, as observed in WMH segmentation [177]. In the unpublished work [177], we developed an adversarial learning-based approach to match the source and target image domains for *cross-scanner* segmentation. We observed that it does not universally work for diverse image acquisitions in the WMH segmentation task. This may be because existing unsupervised domain adaptation techniques (e.g., image-based or feature-based) are application-dependent or shift-dependent. Future research should focus more on understanding domain shift in image space and its impact on network learning.

In a clinical setting, ideally, a user would want a segmentation model to be robust or out-of-box. One example is *Synseg* [178, 179], a contrast-agnostic segmentation model for MR images. As honestly mentioned by the authors, it does not perform well in pathological cases when the model is trained on healthy subjects. Notably, domain shift in a wider setting includes not only contrast variation but also other aspects such as anatomical shift (healthy *vs.* pathological). It is important to note that there is no universal method for settings beyond our assumptions.

When segmentation models fail in deployment, failure analysis [180] can help understand failure behaviors retrospectively and make it possible to *treat* the model by identifying poison (or out-of-distribution) data and attacks in the training set [181]. This step would further improve segmentation robustness in subsequent deployments.

8.2.2 X-to-image synthesis

As demonstrated, image-to-image synthesis successfully augments neuroimages *from almost-invisible to visible* in this dissertation. Some recent alternatives include diffusion models [182,

183] and normalizing flows [184], which are promising for generating high-quality images.

Text-to-image synthesis [185] could be an interesting and promising direction in medical imaging. *DALLE-2*[1] is a recent large-scale text-to-image model for this task. Figure 8.1 shows an example generated by *DALLE-2* with the input caption - 'A high-resolution axial x-ray slice with brain tumor'. Notably, *DALLE-2* is not trained specifically on medical datasets. Why can text-to-image be clinically relevant? From a technical perspective, one straightforward reason is data augmentation and manipulation, e.g., generating tailored samples given some rare but reasonable combinations of text. This could be particularly useful in applications where clinical reports are structured, such as pathology. From a non-professional viewpoint, text is a domain where clinical diagnoses and decisions are recorded based on findings in images and other sources. Hence, building a bridge between text and image might enhance disease understanding and further democratize medical imaging.

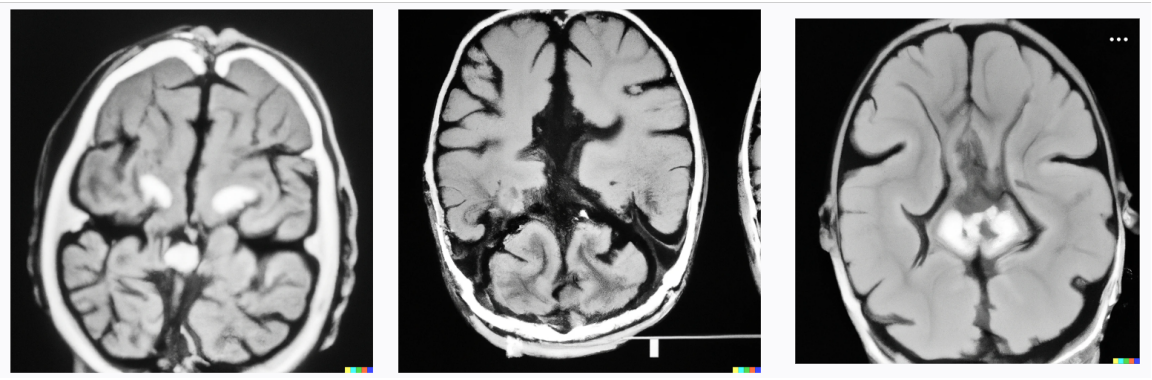


Figure 8.1: 'A high-resolution axial x-ray slice with brain tumor', generated by *DALLE-2*.

It can be argued that the representation of an image should be as compact as possible, i.e., the representation should encode key features essential for *unknown* downstream tasks. While existing works address long-tailed classification problems in self-supervised settings [186, 187], discussions on tackling imbalance issues are generally limited to a single image domain with a uniform distribution. For instance, consider a scenario where the self-supervised model is trained with a set of one million MR images and 100 CT images. The question arises: How many CT characteristics can the representation capture compared to the one trained exclusively on 100 CT images? Or would it be 'overwhelmed' by the attributes of MR images? Since current self-supervised learning frameworks are purely data-driven without any regularization, this might become a significant issue if the goal is for the representation to capture the characteristics of tailed samples in a self-supervised setting.

A key question here is whether there is a benefit to mixing multiple image domains in a self-supervised setting. In medical imaging, the answer could be 'yes' and possibly 'no'. The rationale for 'yes' is the prevailing belief that more training data leads to more effective representations. However, considering MRI as an example, when data from different acquisition settings and medical centers are mixed, it is unclear whether the intra-domain differences (e.g., images from different scanners) would be more pronounced than inter-domain differences (e.g., images from different subtypes) in the learned representations. How can we ensure that self-supervised representations trained on long-tailed datasets remain equally discriminative for different downstream tasks? This aspect seems yet to be explored in medical imaging.

[1] https://openai.com/dall-e-2/

Bibliography

[1] C.-H. Cheng and W.-X. Liu. "Identifying degenerative brain disease using rough set classifier based on wavelet packet method". In: *Journal of clinical medicine* 7.6 (2018), p. 124.

[2] M. E. Wagshul, M. Lucas, K. Ye, M. Izzetoglu, and R. Holtzer. "Multi-modal neuroimaging of dual-task walking: Structural MRI and fNIRS analysis reveals prefrontal grey matter volume moderation of brain activation in older adults". In: *Neuroimage* 189 (2019), pp. 745–754.

[3] A. Mohamed, E. I. Zacharaki, D. Shen, and C. Davatzikos. "Deformable registration of brain tumor images via a statistical model of tumor-induced deformation". In: *Medical image analysis* 10.5 (2006), pp. 752–763.

[4] M. Liu, D. Zhang, and D. Shen. "Relationship induced multi-template learning for diagnosis of Alzheimer's disease and mild cognitive impairment". In: *IEEE transactions on medical imaging* 35.6 (2016), pp. 1463–1474.

[5] J. Ouyang, E. Adeli, K. M. Pohl, Q. Zhao, and G. Zaharchuk. "Representation disentanglement for multi-modal brain mri analysis". In: *International Conference on Information Processing in Medical Imaging*. Springer. 2021, pp. 321–333.

[6] L. K. Tremblay, C. Hammill, S. H. Ameis, M. Bhaijiwala, D. J. Mabbott, E. Anagnostou, J. P. Lerch, and R. J. Schachar. "Tracking inhibitory control in youth with ADHD: A multi-modal neuroimaging approach". In: *Frontiers in Psychiatry* 11 (2020), p. 00831.

[7] A. M. Rauschecker, J. D. Rudie, L. Xie, J. Wang, M. T. Duong, E. J. Botzolakis, A. M. Kovalovich, J. Egan, T. C. Cook, R. N. Bryan, et al. "Artificial intelligence system approaching neuroradiologist-level differential diagnosis accuracy at brain MRI". In: *Radiology* 295.3 (2020), p. 626.

[8] R. Shaul, I. David, O. Shitrit, and T. R. Raviv. "Subsampled brain MRI reconstruction by generative adversarial neural networks". In: *Medical Image Analysis* 65 (2020), p. 101747.

[9] A. Sanaat, H. Arabi, I. Mainta, V. Garibotto, and H. Zaidi. "Projection space implementation of deep learning–guided low-dose brain PET imaging improves performance over implementation in image space". In: *Journal of Nuclear Medicine* 61.9 (2020), pp. 1388–1396.

[10] C. N. Ladefoged, A. E. Hansen, O. M. Henriksen, F. J. Bruun, L. Eikenes, S. K. Øen, A. Karlberg, L. Højgaard, I. Law, and F. L. Andersen. "AI-driven attenuation correction for brain PET/MRI: Clinical evaluation of a dementia cohort and importance of the training group size". In: *Neuroimage* 222 (2020), p. 117221.

[11] E. Pace and M. Borg. "Optimisation of a paediatric CT brain protocol: a figure-of-merit approach". In: *Radiation Protection Dosimetry* 182.3 (2018), pp. 394–404.

[12] L. Evans, M. Fitzgerald, D. Varma, and B. Mitra. "A novel approach to improving the interpretation of CT brain in trauma". In: *Injury* 49.1 (2018), pp. 56–61.

[13] P. Anbeek, K. L. Vincken, G. S. Van Bochove, M. J. Van Osch, and J. van der Grond. "Probabilistic segmentation of brain tissue in MR imaging". In: *Neuroimage* 27.4 (2005), pp. 795–804.

[14] P. Eichinger, A. Hock, S. Schön, C. Preibisch, J. S. Kirschke, M. Mühlau, C. Zimmer, and B. Wiestler. "Acceleration of double inversion recovery sequences in multiple sclerosis with compressed sensing". In: *Investigative Radiology* 54.6 (2019), pp. 319–324.

[15] A. S. Nielsen, R. P. Kinkel, E. Tinelli, T. Benner, J. Cohen-Adad, and C. Mainero. "Focal cortical lesion detection in multiple sclerosis: 3 Tesla DIR versus 7 Tesla FLASH-T2". In: *Journal of Magnetic Resonance Imaging* 35.3 (2012), pp. 537–542.

[16] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, et al. "The multimodal brain tumor image segmentation benchmark (BRATS)". In: *IEEE transactions on medical imaging* 34.10 (2014), pp. 1993–2024.

[17] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, R. T. Shinohara, C. Berger, S. M. Ha, M. Rozycki, et al. "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge". In: *arXiv preprint arXiv:1811.02629* (2018).

[18] G. Hinton. "Deep learning—a technology with the potential to transform health care". In: *Jama* 320.11 (2018), pp. 1101–1102.

[19] S. M. Plis, D. R. Hjelm, R. Salakhutdinov, E. A. Allen, H. J. Bockholt, J. D. Long, H. J. Johnson, J. S. Paulsen, J. A. Turner, and V. D. Calhoun. "Deep learning for neuroimaging: a validation study". In: *Frontiers in neuroscience* 8 (2014), p. 229.

[20] H.-I. Suk, S.-W. Lee, D. Shen, A. D. N. Initiative, et al. "Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis". In: *NeuroImage* 101 (2014), pp. 569–582.

[21] D. C. Castro, I. Walker, and B. Glocker. "Causality matters in medical imaging". In: *Nature Communications* 11.1 (2020), pp. 1–10.

[22] A. S. Lundervold and A. Lundervold. "An overview of deep learning in medical imaging focusing on MRI". In: *Zeitschrift für Medizinische Physik* 29.2 (2019), pp. 102–127.

[23] A. Elnakib, G. Gimel'farb, J. S. Suri, and A. El-Baz. "Medical image segmentation: a brief survey". In: *Multi Modality State-of-the-Art Medical Image Segmentation and Registration Methodologies*. Springer, 2011, pp. 1–39.

[24] L. Wang, G. Li, F. Shi, X. Cao, C. Lian, D. Nie, M. Liu, H. Zhang, G. Li, Z. Wu, et al. "Volume-based analysis of 6-month-old infant brain MRI for autism biomarker identification and early diagnosis". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pp. 411–419.

[25] P. F. Christ, M. E. A. Elshaer, F. Ettlinger, S. Tatavarty, M. Bickel, P. Bilic, M. Rempfler, M. Armbruster, F. Hofmann, M. D'Anastasi, et al. "Automatic liver and lesion segmentation in CT using cascaded fully convolutional neural networks and 3D conditional random fields". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2016, pp. 415–423.

[26] O. Maier, B. H. Menze, J. von der Gablentz, L. Häni, M. P. Heinrich, M. Liebrand, S. Winzeck, A. Basit, P. Bentley, L. Chen, et al. "ISLES 2015-A public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI". In: *Medical image analysis* 35 (2017), pp. 250–269.

[27] **H. Li**, G. Jiang, J. Zhang, R. Wang, Z. Wang, W.-S. Zheng, and B. Menze. "Fully convolutional network ensembles for white matter hyperintensities segmentation in MR images". In: ***NeuroImage*** 183 (2018), pp. 650–665.

[28] M. Sezgin and B. Sankur. "Survey over image thresholding techniques and quantitative performance evaluation". In: *Journal of Electronic imaging* 13.1 (2004), pp. 146–166.

[29] J. Franklin. "The elements of statistical learning: data mining, inference and prediction". In: *The Mathematical Intelligencer* 27.2 (2005), pp. 83–85.

[30] D. W. Scott. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 2015.

[31] D. Smeets, B. Stijnen, D. Loeckx, B. De Dobbelaer, and P. Suetens. "Segmentation of liver metastases using a level set method with spiral-scanning technique and supervised fuzzy pixel classification". In: *MICCAI workshop*. Vol. 42. 2008, p. 43.

[32] D. Jiménez Carretero, L. Fernández de Manuel, J. Pascau González Garzón, J. M. Tellado, E. Ramon, M. Desco Menéndez, A. Santos, and M. J. Ledesma Carbayo. "Optimal multiresolution 3D level-set method for liver segmentation incorporating local curvature constraints". In: (2011).

[33] H. Wang, J. W. Suh, S. R. Das, J. B. Pluta, C. Craige, and P. A. Yushkevich. "Multi-atlas segmentation with joint label fusion". In: *IEEE transactions on pattern analysis and machine intelligence* 35.3 (2013), pp. 611–623.

[34] L. M. Koch, M. Rajchl, W. Bai, C. F. Baumgartner, T. Tong, J. Passerat-Palmbach, P. Aljabar, and D. Rueckert. "Multi-atlas segmentation using partially annotated data: methods and annotation strategies". In: *IEEE transactions on pattern analysis and machine intelligence* 40.7 (2017), pp. 1683–1696.

[35] L. Massoptier and S. Casciaro. "A new fully automatic and robust algorithm for fast segmentation of liver tissue and tumors from CT scans". In: *European radiology* 18.8 (2008), p. 1658.

[36] H. Frigui and R. Krishnapuram. "A robust competitive clustering algorithm with applications in computer vision". In: *Ieee transactions on pattern analysis and machine intelligence* 21.5 (1999), pp. 450–465.

[37] Y. Häme. "Liver tumor segmentation using implicit surface evolution". In: *The Midas Journal* (2008), pp. 1–10.

[38] M. Sonka, V. Hlavac, and R. Boyle. *Image processing, analysis, and machine vision*. Cengage Learning, 2014.

[39] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid. "Local features and kernels for classification of texture and object categories: A comprehensive study". In: *International journal of computer vision* 73.2 (2007), pp. 213–238.

[40] R. M. Haralick. "Statistical and structural approaches to texture". In: *Proceedings of the IEEE* 67.5 (1979), pp. 786–804.

[41] M. Yang, Y. Yuan, X. Li, and P. Yan. "Medical Image Segmentation Using Descriptive Image Features." In: *BMVC*. Citeseer. 2011, pp. 1–11.

[42] C.-C. Chang and C.-J. Lin. "LIBSVM: A library for support vector machines". In: *ACM transactions on intelligent systems and technology (TIST)* 2.3 (2011), pp. 1–27.

[43] E. Geremia, B. H. Menze, O. Clatz, E. Konukoglu, A. Criminisi, and N. Ayache. "Spatial decision forests for MS lesion segmentation in multi-channel MR images". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2010, pp. 111–118.

[44] M. P. Heinrich and M. Blendowski. "Multi-organ segmentation using vantage point forests and binary context features". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2016, pp. 598–606.

[45] T. Tong, R. Wolz, P. Coupé, J. V. Hajnal, D. Rueckert, A. D. N. Initiative, et al. "Segmentation of MR images via discriminative dictionary learning and sparse coding: Application to hippocampus labeling". In: *NeuroImage* 76 (2013), pp. 11–23.

[46] K. Kamnitsas, C. Ledig, V. F. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker. "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation". In: *Medical image analysis* 36 (2017), pp. 61–78.

[47] J. Long, E. Shelhamer, and T. Darrell. "Fully convolutional networks for semantic segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3431–3440.

[48] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang. "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation". In: *IEEE transactions on medical imaging* 39.6 (2019), pp. 1856–1867.

[49] X. Yan, W. Jiang, Y. Shi, and C. Zhuo. "Ms-nas: Multi-scale neural architecture search for medical image segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2020, pp. 388–397.

[50] F. Isensee, P. F. Jaeger, S. 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 (2021), pp. 203–211.

[51] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. "Focal loss for dense object detection". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2980–2988.

[52] F. Milletari, N. Navab, and S.-A. Ahmadi. "V-net: Fully convolutional neural networks for volumetric medical image segmentation". In: *2016 fourth international conference on 3D vision (3DV)*. IEEE. 2016, pp. 565–571.

[53] S. S. M. Salehi, D. Erdogmus, and A. Gholipour. "Tversky loss function for image segmentation using 3D fully convolutional deep networks". In: *International workshop on machine learning in medical imaging*. Springer. 2017, pp. 379–387.

[54] D. Karimi and S. E. Salcudean. "Reducing the hausdorff distance in medical image segmentation with convolutional neural networks". In: *IEEE Transactions on medical imaging* 39.2 (2019), pp. 499–513.

[55] Y. Liu, H. Chen, Y. Chen, W. Yin, and C. Shen. "Generic perceptual loss for modeling structured output dependencies". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 5424–5432.

[56] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. "How transferable are features in deep neural networks?" In: *Advances in neural information processing systems* 27 (2014).

[57] C. S. Perone, P. Ballester, R. C. Barros, and J. Cohen-Adad. "Unsupervised domain adaptation for medical imaging segmentation with self-ensembling". In: *NeuroImage* 194 (2019), pp. 1–11.

[58] K. Kamnitsas, C. Baumgartner, C. Ledig, V. Newcombe, J. Simpson, A. Kane, D. Menon, A. Nori, A. Criminisi, D. Rueckert, et al. "Unsupervised domain adaptation in brain lesion segmentation with adversarial networks". In: *International conference on information processing in medical imaging*. Springer. 2017, pp. 597–609.

[59] N. Karani, E. Erdil, K. Chaitanya, and E. Konukoglu. "Test-time adaptable neural networks for robust medical image segmentation". In: *Medical Image Analysis* 68 (2021), p. 101907.

[60] Q. Dou, C. Ouyang, C. Chen, H. Chen, and P.-A. Heng. "Unsupervised cross-modality domain adaptation of convnets for biomedical image segmentations with adversarial loss". In: *arXiv preprint arXiv:1804.10916* (2018).

[61] 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 AAAI conference on artificial intelligence*. Vol. 33. 01. 2019, pp. 865–872.

[62] C. Ouyang, K. Kamnitsas, C. Biffi, J. Duan, and D. Rueckert. "Data efficient unsupervised domain adaptation for cross-modality image segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pp. 669–677.

[63] C. Ouyang, C. Chen, S. Li, Z. Li, C. Qin, W. Bai, and D. Rueckert. "Causality-inspired Single-source Domain Generalization for Medical Image Segmentation". In: *arXiv preprint arXiv:2111.12525* (2021).

[64] C. Chen, Z. Li, C. Ouyang, M. Sinclair, W. Bai, and D. Rueckert. "MaxStyle: Adversarial Style Composition for Robust Medical Image Segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2022, pp. 151–161.

[65] S. Ben Atitallah, M. Driss, W. Boulila, and H. Ben Ghezala. "Randomly initialized convolutional neural network for the recognition of COVID-19 using X-ray images". In: *International journal of imaging systems and technology* 32.1 (2022), pp. 55–73.

[66] V. Ramanujan, M. Wortsman, A. Kembhavi, A. Farhadi, and M. Rastegari. "What's hidden in a randomly weighted neural network?" In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 11893–11902.

[67] X. Glorot and Y. Bengio. "Understanding the difficulty of training deep feedforward neural networks". In: *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings. 2010, pp. 249–256.

[68] Z. Zhou, V. Sodha, M. M. Rahman Siddiquee, R. Feng, N. Tajbakhsh, M. B. Gotway, and J. Liang. "Models genesis: Generic autodidactic models for 3d medical image analysis". In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2019, pp. 384–393.

[69] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. "Imagenet large scale visual recognition challenge". In: *International journal of computer vision* 115.3 (2015), pp. 211–252.

[70] K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu. "Contrastive learning of global and local features for medical image segmentation with limited annotations". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 12546–12558.

[71] R. Bitar, G. Leung, R. Perng, S. Tadros, A. R. Moody, J. Sarrazin, C. McGregor, M. Christakis, S. Symons, A. Nelson, et al. "MR pulse sequences: what every radiologist wants to know but is afraid to ask". In: *Radiographics* 26.2 (2006), pp. 513–537.

[72] P. Schmidt, V. Pongratz, P. Küster, D. Meier, J. Wuerfel, C. Lukas, B. Bellenberg, F. Zipp, S. Groppa, P. G. Sämann, et al. "Automated segmentation of changes in FLAIR-hyperintense white matter lesions in multiple sclerosis on serial magnetic resonance imaging". In: *NeuroImage: Clinical* 23 (2019), p. 101849.

[73] J. Liu, M. Kocak, M. Supanich, and J. Deng. "Motion artifacts reduction in brain MRI by means of a deep residual network with densely connected multi-resolution blocks (DRN-DCMB)". In: *Magnetic resonance imaging* 71 (2020), pp. 69–79.

[74] Y. Huang, L. Shao, and A. F. Frangi. "Cross-modality image synthesis via weakly coupled and geometry co-regularized joint dictionary learning". In: *IEEE transactions on medical imaging* 37.3 (2017), pp. 815–827.

[75] S. Roy, A. Carass, and J. L. Prince. "Magnetic resonance image example-based contrast synthesis". In: *IEEE transactions on medical imaging* 32.12 (2013), pp. 2348–2363.

[76] A. Jog, S. Roy, A. Carass, and J. L. Prince. "Magnetic resonance image synthesis through patch regression". In: *2013 IEEE 10th International Symposium on Biomedical Imaging*. IEEE. 2013, pp. 350–353.

[77] R. Li, W. Zhang, H.-I. Suk, L. Wang, J. Li, D. Shen, and S. Ji. "Deep learning based imaging data completion for improved brain disease diagnosis". In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2014, pp. 305–312.

[78] S. U. Dar, M. Yurt, L. Karacan, A. Erdem, E. Erdem, and T. Cukur. "Image synthesis in multi-contrast MRI with conditional generative adversarial networks". In: *IEEE transactions on medical imaging* 38.10 (2019), pp. 2375–2388.

[79] L. Qu, S. Wang, P.-T. Yap, and D. Shen. "Wavelet-based semi-supervised adversarial learning for synthesizing realistic 7T from 3T MRI". In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2019, pp. 786–794.

[80] A. Jog, A. Carass, S. Roy, D. L. Pham, and J. L. Prince. "Random forest regression for magnetic resonance image synthesis". In: *Medical image analysis* 35 (2017), pp. 475–488.

[81] A. Jog, A. Carass, D. L. Pham, and J. L. Prince. "Random forest flair reconstruction from t 1, t 2, and p d-weighted mri". In: *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2014, pp. 1079–1082.

[82] A. Chartsias, T. Joyce, M. V. Giuffrida, and S. A. Tsaftaris. "Multimodal MR synthesis via modality-invariant latent representation". In: *IEEE transactions on medical imaging* 37.3 (2017), pp. 803–814.

[83] T. Joyce, A. Chartsias, and S. A. Tsaftaris. "Robust multi-modal MR image synthesis". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2017, pp. 347–355.

[84] W. Wei, E. Poirion, B. Bodini, S. Durrleman, O. Colliot, B. Stankoff, and N. Ayache. "FLAIR MR image synthesis by using 3D fully convolutional networks for multiple sclerosis". In: *ISMRM-ESMRMB 2018-Joint Annual Meeting*. 2018, pp. 1–6.

[85] M. Yurt, S. U. Dar, A. Erdem, E. Erdem, K. K. Oguz, and T. Çukur. "mustGAN: multi-stream generative adversarial networks for MR image synthesis". In: *Medical image analysis* 70 (2021), p. 101944.

[86] C. C. Jaffe et al. "Measures of response: RECIST, WHO, and new alternatives". In: *J Clin Oncol* 24.20 (2006), pp. 3245–51.

[87] V. Kumar, Y. Gu, S. Basu, A. Berglund, S. A. Eschrich, M. B. Schabath, K. Forster, H. J. Aerts, A. Dekker, D. Fenstermacher, et al. "Radiomics: the process and the challenges". In: *Magnetic resonance imaging* 30.9 (2012), pp. 1234–1248.

[88] O. Holub and S. T. Ferreira. "Quantitative histogram analysis of images". In: *Computer physics communications* 175.9 (2006), pp. 620–623.

[89] I. El Naqa, P. Grigsby, A. Apte, E. Kidd, E. Donnelly, D. Khullar, S. Chaudhari, D. Yang, M. Schmitt, R. Laforest, et al. "Exploring feature-based approaches in PET images for predicting cancer treatment outcomes". In: *Pattern recognition* 42.6 (2009), pp. 1162–1171.

[90] F. O'sullivan, S. Roy, J. O'sullivan, C. Vernon, and J. Eary. "Incorporation of tumor shape into an assessment of spatial heterogeneity for human sarcomas imaged with FDG-PET". In: *Biostatistics* 6.2 (2005), pp. 293–301.

[91] R. C. Gonzalez. *Digital image processing*. Pearson education india, 2009.

[92] G. Castellano, L. Bonilha, L. Li, and F. Cendes. "Texture analysis of medical images". In: *Clinical radiology* 59.12 (2004), pp. 1061–1069.

[93] J. Zhang and T. Tan. "Brief review of invariant texture analysis methods". In: *Pattern recognition* 35.3 (2002), pp. 735–747.

[94] W. Li, M. Coats, J. Zhang, and S. J. McKenna. "Discriminating dysplasia: Optical tomographic texture analysis of colorectal polyps". In: *Medical image analysis* 26.1 (2015), pp. 57–69.

[95] S. Manivannan, W. Li, S. Akbar, R. Wang, J. Zhang, and S. J. McKenna. "An automated pattern recognition system for classifying indirect immunofluorescence images of HEp-2 cells and specimens". In: *Pattern Recognition* 51 (2016), pp. 12–26.

[96] G. Zhao, T. Ahonen, J. Matas, and M. Pietikainen. "Rotation-invariant image and video description with local binary pattern features". In: *IEEE transactions on image processing* 21.4 (2011), pp. 1465–1477.

[97] F. Bianconi and A. Fernández. "Evaluation of the effects of Gabor filter parameters on texture classification". In: *Pattern recognition* 40.12 (2007), pp. 3325–3335.

[98] Y. Zhang, R. Jin, and Z.-H. Zhou. "Understanding bag-of-words model: a statistical framework". In: *International journal of machine learning and cybernetics* 1.1 (2010), pp. 43–52.

[99] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek. "Image classification with the fisher vector: Theory and practice". In: *International journal of computer vision* 105.3 (2013), pp. 222–245.

[100] R. Paul, S. H. Hawkins, Y. Balagurunathan, M. Schabath, R. J. Gillies, L. O. Hall, and D. B. Goldgof. "Deep feature transfer learning in combination with traditional features predicts survival among patients with lung adenocarcinoma". In: *Tomography* 2.4 (2016), pp. 388–395.

[101] W. Shen, M. Zhou, F. Yang, D. Yu, D. Dong, C. Yang, Y. Zang, and J. Tian. "Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification". In: *Pattern Recognition* 61 (2017), pp. 663–673.

[102] B. Q. Huynh, H. Li, and M. L. Giger. "Digital mammographic tumor classification using transfer learning from deep convolutional neural networks". In: *Journal of Medical Imaging* 3.3 (2016), p. 034501.

[103] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. "Dermatologist-level classification of skin cancer with deep neural networks". In: *nature* 542.7639 (2017), pp. 115–118.

[104] L. Hirsch, Y. Huang, S. Luo, C. Rossi Saccarelli, R. Lo Gullo, I. Daimiel Naranjo, A. G. Bitencourt, N. Onishi, E. S. Ko, D. Leithner, et al. "Radiologist-level performance by using deep learning for segmentation of breast cancers on MRI scans". In: *Radiology: Artificial Intelligence* 4.1 (2021), e200231.

[105] A. Cadrin-Chênevert. "Moving from ImageNet to RadImageNet for improved transfer learning and generalizability". In: *Radiology: Artificial Intelligence* 4.5 (2022), e220126.

[106] Z. Liu, S. Wang, D. Dong, J. Wei, C. Fang, X. Zhou, K. Sun, L. Li, B. Li, M. Wang, et al. "The applications of radiomics in precision diagnosis and treatment of oncology: opportunities and challenges". In: *Theranostics* 9.5 (2019), p. 1303.

[107] L. Pantoni. "Cerebral small vessel disease: from pathogenesis and clinical characteristics to therapeutic challenges". In: *The Lancet Neurology* 9.7 (2010), pp. 689–701.

[108] K. W. Kim, J. R. MacFall, and M. E. Payne. "Classification of white matter lesions on magnetic resonance imaging in elderly persons". In: *Biological psychiatry* 64.4 (2008), pp. 273–280.

[109] S. Debette and H. Markus. "The clinical importance of white matter hyperintensities on brain magnetic resonance imaging: systematic review and meta-analysis". In: *Bmj* 341 (2010).

[110] J. Grimaud, M. Lai, J. Thorpe, P. Adeleine, L. Wang, G. Barker, D. Plummer, P. Tofts, W. McDonald, and D. Miller. "Quantification of MRI lesion load in multiple sclerosis: a comparison of three computer-assisted techniques". In: *Magnetic resonance imaging* 14.5 (1996), pp. 495–505.

[111] H. J. Kuijf, J. M. Biesbroek, J. De Bresser, R. Heinen, S. Andermatt, M. Bento, M. Berseth, M. Belyaev, M. J. Cardoso, A. Casamitjana, **H. Li**, et al. "Standardized assessment of automatic segmentation of white matter hyperintensities and results of the WMH segmentation challenge". In: *IEEE transactions on medical imaging (**TMI**)* 38.11 (2019), pp. 2556–2568.

[112] B. N. Mathur. "The claustrum in review". In: *Frontiers in systems neuroscience* 8 (2014), p. 48.

[113] C. M. Torgerson, A. Irimia, S. M. Goh, and J. D. Van Horn. "The DTI connectivity of the human claustrum". In: *Human brain mapping* 36.3 (2015), pp. 827–838.

[114] B. Zingg, H. Hintiryan, L. Gou, M. Y. Song, M. Bay, M. S. Bienkowski, N. N. Foster, S. Yamashita, I. Bowman, A. W. Toga, et al. "Neural networks of the mouse neocortex". In: *Cell* 156.5 (2014), pp. 1096–1111.

[115] S. P. Brown, B. N. Mathur, S. R. Olsen, P.-H. Luppi, M. E. Bickford, and A. Citri. "New breakthroughs in understanding the role of functional interactions between the neocortex and the claustrum". In: *Journal of Neuroscience* 37.45 (2017), pp. 10877–10881.

[116] R. Remedios, N. K. Logothetis, and C. Kayser. "A role of the claustrum in auditory scene analysis by reflecting sensory change". In: *Frontiers in systems neuroscience* 8 (2014), p. 44.

[117] H. Bruguier, R. Suarez, P. Manger, A. Hoerder-Suabedissen, A. M. Shelton, D. K. Oliver, A. M. Packer, J. L. Ferran, F. García-Moreno, L. Puelles, et al. "In search of common developmental and evolutionary origin of the claustrum and subplate". In: *Journal of Comparative Neurology* 528.17 (2020), pp. 2956–2977.

[118] C. Watson and L. Puelles. "Developmental gene expression in the mouse clarifies the organization of the claustrum and related endopiriform nuclei". In: *Journal of Comparative Neurology* 525.6 (2017), pp. 1499–1508.

[119] A. Arrigo, E. Mormina, A. Calamuneri, M. Gaeta, F. Granata, S. Marino, G. Anastasi, D. Milardi, and A. Quartarone. "Inter-hemispheric claustral connections in human brain: a constrained spherical deconvolution-based study". In: *Clinical neuroradiology* 27.3 (2017), pp. 275–281.

[120] D. Milardi, P. Bramanti, C. Milazzo, G. Finocchio, A. Arrigo, G. Santoro, F. Trimarchi, A. Quartarone, G. Anastasi, and M. Gaeta. "Cortical and subcortical connections of the human claustrum revealed in vivo by constrained spherical deconvolution tractography". In: *Cerebral Cortex* 25.2 (2015), pp. 406–414.

[121] O. Ronneberger, P. Fischer, and T. Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, pp. 234–241.

[122] J. Smythies, L. Edelstein, and V. S. Ramachandran. *The claustrum: structural, functional, and clinical neuroscience*. Academic Press, 2013.

[123] P. J. Brittain, S. F. Walsh, K.-W. Nam, V. Giampietro, V. Karolis, R. M. Murray, S. Bhattacharyya, A. Kalpakidou, and C. Nosarti. "Neural compensation in adulthood following very preterm birth demonstrated during a visual paired associates learning task". In: *NeuroImage: Clinical* 6 (2014), pp. 54–63.

[124] S. Berman, R. Schurr, G. Atlan, A. Citri, and A. A. Mezer. "Automatic segmentation of the dorsal claustrum in humans using in vivo high-resolution MRI". In: *Cerebral Cortex Communications* 1.1 (2020), tgaa062.

[125] A. A. Albishri, S. J. H. Shah, A. Schmiedler, S. S. Kang, and Y. Lee. "Automated human claustrum segmentation using deep learning technologies". In: *arXiv preprint arXiv:1911.07515* (2019).

[126] **H. Li**, A. Menegaux, B. Schmitz-Koep, A. Neubauer, F. J. Bäuerlein, S. Shit, C. Sorg, B. Menze, and D. Hedderich. "Automated claustrum segmentation in human brain MRI using deep learning". In: ***Human Brain Mapping*** 42.18 (2021), pp. 5862–5872.

[127] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. "Unpaired image-to-image translation using cycle-consistent adversarial networks". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2223–2232.

[128] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. "Image quality assessment: from error visibility to structural similarity". In: *IEEE transactions on image processing* 13.4 (2004), pp. 600–612.

[129] H. Feng, J. Cao, H. Wang, Y. Xie, D. Yang, J. Feng, and B. Chen. "A knowledge-driven feature learning and integration method for breast cancer diagnosis on multi-sequence MRI". In: *Magnetic resonance imaging* 69 (2020), pp. 40–48.

[130] D. Avola, L. Cinque, A. Fagioli, S. Filetti, G. Grani, and E. Rodolà. "Multimodal feature fusion and knowledge-driven learning via experts consult for thyroid nodule classification". In: *IEEE Transactions on Circuits and Systems for Video Technology* 32.5 (2021), pp. 2527–2534.

[131] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. "A simple framework for contrastive learning of visual representations". In: *International conference on machine learning*. PMLR. 2020, pp. 1597–1607.

[132] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. "Momentum contrast for unsupervised visual representation learning". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 9729–9738.

[133] A. v. d. Oord, Y. Li, and O. Vinyals. "Representation learning with contrastive predictive coding". In: *arXiv preprint arXiv:1807.03748* (2018).

[134] A. Krizhevsky, G. Hinton, et al. "Learning multiple layers of features from tiny images". In: (2009).

[135] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio. *Deep learning*. Vol. 1. MIT press Cambridge, 2016.

[136] Y. LeCun, Y. Bengio, and G. Hinton. "Deep learning". In: *nature* 521.7553 (2015), p. 436.

[137] A. Burkov. *The hundred-page machine learning book*. Vol. 1. Andriy Burkov Quebec City, QC, Canada, 2019.

[138] K. P. Murphy. *Probabilistic machine learning: an introduction*. MIT press, 2022.

[139] Y. Zhao. "Deep learning based medical image segmentation and classification for artificial intelligence healthcare". PhD thesis. Technische Universität München, 2021.

[140] X. Glorot, A. Bordes, and Y. Bengio. "Deep sparse rectifier neural networks". In: *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings. 2011, pp. 315–323.

[141] B. Xu, N. Wang, T. Chen, and M. Li. "Empirical evaluation of rectified activations in convolutional network". In: *arXiv preprint arXiv:1505.00853* (2015).

[142] I. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio. "Maxout networks". In: *International conference on machine learning*. PMLR. 2013, pp. 1319–1327.

[143] P. Ramachandran, B. Zoph, and Q. V. Le. "Swish: a self-gated activation function". In: *arXiv preprint arXiv:1710.05941* 7.1 (2017), p. 5.

[144] A. Maier, C. Syben, T. Lasser, and C. Riess. "A gentle introduction to deep learning in medical image processing". In: *Zeitschrift für Medizinische Physik* 29.2 (2019), pp. 86–101.

[145] Y. Bengio, A. Courville, and P. Vincent. "Representation learning: A review and new perspectives". In: *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013), pp. 1798–1828.

[146] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. "Learning representations by back-propagating errors". In: *nature* 323.6088 (1986), pp. 533–536.

[147] D. Shen, G. Wu, and H.-I. Suk. "Deep learning in medical image analysis". In: *Annual review of biomedical engineering* 19 (2017), pp. 221–248.

[148] A. Krizhevsky, I. Sutskever, and G. E. Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems*. 2012, pp. 1097–1105.

[149] K. Simonyan and A. Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014).

[150] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. "Going deeper with convolutions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9.

[151] K. He, X. Zhang, S. Ren, and J. Sun. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[152] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. "Densely connected convolutional networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4700–4708.

[153] M. Tan and Q. Le. "Efficientnet: Rethinking model scaling for convolutional neural networks". In: *International conference on machine learning*. PMLR. 2019, pp. 6105–6114.

[154] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger. "3D U-Net: learning dense volumetric segmentation from sparse annotation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2016, pp. 424–432.

[155] G. Lin, A. Milan, C. Shen, and I. Reid. "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1925–1934.

[156] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng. "H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes". In: *IEEE transactions on medical imaging* 37.12 (2018), pp. 2663–2674.

[157] S. Guan, A. A. Khan, S. Sikdar, and P. V. Chitnis. "Fully Dense UNet for 2-D Sparse Photoacoustic Tomography Artifact Removal". In: *IEEE journal of biomedical and health informatics* 24.2 (2019), pp. 568–576.

[158] F. Yu and V. Koltun. "Multi-scale context aggregation by dilated convolutions". In: *arXiv preprint arXiv:1511.07122* (2015).

[159] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei. "Deformable convolutional networks". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 764–773.

[160] R. Zhang, F. Zhu, J. Liu, and G. Liu. "Depth-wise separable convolutions and multi-level pooling for an efficient spatial CNN-based steganalysis". In: *IEEE Transactions on Information Forensics and Security* 15 (2019), pp. 1138–1150.

[161] M. D. Zeiler and R. Fergus. "Visualizing and understanding convolutional networks". In: *European conference on computer vision*. Springer. 2014, pp. 818–833.

[162] V. Dumoulin and F. Visin. "A guide to convolution arithmetic for deep learning". In: *arXiv preprint arXiv:1603.07285* (2016).

[163] S. Jadon. "A survey of loss functions for semantic segmentation". In: *arXiv preprint arXiv:2006.14822* (2020).

[164] H. Kervadec, J. Bouchtiba, C. Desrosiers, E. Granger, J. Dolz, and I. B. Ayed. "Boundary loss for highly unbalanced segmentation". In: *International conference on medical imaging with deep learning*. 2019, pp. 285–296.

[165] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. "Generative adversarial networks". In: *Communications of the ACM* 63.11 (2020), pp. 139–144.

[166] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio. "Generative Adversarial Nets". In: *NIPS*. 2014.

[167] I. Goodfellow. "Nips 2016 tutorial: Generative adversarial networks". In: *arXiv preprint arXiv:1701.00160* (2016).

[168] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers. "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning". In: *IEEE transactions on medical imaging* 35.5 (2016), pp. 1285–1298.

[169] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. "Adversarial discriminative domain adaptation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 7167–7176.

[170] A. Neubauer*, **H. Li***, J. Wendt, B. Schmitz-Koep, A. Menegaux, D. Schinz, B. Menze, C. Zimmer, C. Sorg, and D. M. Hedderich. "Efficient claustrum segmentation in T2-weighted neonatal brain MRI using transfer learning from adult scans". In: ***Clinical Neuroradiology*** (2022), pp. 1–12.

[171] **H. Li***, J. C. Paetzold*, A. Sekuboyina, F. Kofler, J. Zhang, J. S. Kirschke, B. Wiestler, and B. Menze. "DiamondGAN: unified multi-modal generative adversarial networks for MRI sequences synthesis". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (**MICCAI**)*. Springer. 2019, pp. 795–803.

[172] T. Finck*, **H. Li***, L. Grundl, P. Eichinger, M. Bussas, M. Mühlau, B. Menze, and B. Wiestler. "Deep-learning generated synthetic double inversion recovery images improve multiple sclerosis lesion detection". In: ***Investigative Radiology*** 55.5 (2020), pp. 318–323.

[173] T. Finck*, **H. Li***, S. Schlaeger, L. Grundl, N. Sollmann, B. Bender, E. Bürkle, C. Zimmer, J. Kirschke, B. Menze, et al. "Uncertainty-aware and lesion-specific image synthesis in multiple sclerosis magnetic resonance imaging: a multicentric validation study". In: ***Frontiers in Neuroscience*** 16 (2022).

[174] Q. Hu*, **H. Li***, and J. Zhang. "Domain-adaptive 3D medical image synthesis: an efficient unsupervised approach". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (**MICCAI**)* (2022), pp. 495–504.

[175] **H. Li**, F.-F. Xue, K. Chaitanya, S. Luo, I. Ezhov, B. Wiestler, J. Zhang, and B. Menze. "Imbalance-aware self-supervised learning for 3d radiomic representations". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (**MICCAI**)*. Springer. 2021, pp. 36–46.

[176] M. Ghafoorian, A. Mehrtash, T. Kapur, N. Karssemeijer, E. Marchiori, M. Pesteie, C. R. Guttmann, F.-E. d. Leeuw, C. M. Tempany, B. v. Ginneken, et al. "Transfer learning for domain adaptation in MRI: Application in brain lesion segmentation". In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2017, pp. 516–524.

[177] H. Li, T. Loehr, A. Sekuboyina, J. Zhang, B. Wiestler, and B. Menze. "Domain adaptive medical image segmentation via adversarial learning of disease-specific spatial patterns". In: *arXiv preprint arXiv:2001.09313* (2020).

[178] B. Billot, D. Greve, K. Van Leemput, B. Fischl, J. E. Iglesias, and A. V. Dalca. "A learning strategy for contrast-agnostic MRI segmentation". In: *arXiv preprint arXiv:2003.01995* (2020).

[179] B. Billot, D. N. Greve, O. Puonti, A. Thielscher, K. Van Leemput, B. Fischl, A. V. Dalca, and J. E. Iglesias. "Synthseg: Domain randomisation for segmentation of brain mri scans of any contrast and resolution". In: *arXiv preprint arXiv:2107.09559* (2021).

[180] T. Henn, Y. Sakamoto, C. Jacquet, S. Yoshizawa, M. Andou, S. Tchen, R. Saga, H. Ishihara, K. Shimizu, Y. Li, et al. "A Principled Approach to Failure Analysis and Model Repairment: Demonstration in Medical Imaging". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2021, pp. 509–518.

[181] R. Tanno, M. F. Pradier, A. Nori, and Y. Li. "Repairing Neural Networks by Leaving the Right Past Behind". In: *arXiv preprint arXiv:2207.04806* (2022).

[182] W. H. Pinaya, P.-D. Tudosiu, J. Dafflon, P. F. Da Costa, V. Fernandez, P. Nachev, S. Ourselin, and M. J. Cardoso. "Brain imaging generation with latent diffusion models". In: *MICCAI Workshop on Deep Generative Models*. Springer. 2022, pp. 117–126.

[183] P. Dhariwal and A. Nichol. "Diffusion models beat gans on image synthesis". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 8780–8794.

[184] G. Yang, X. Huang, Z. Hao, M.-Y. Liu, S. Belongie, and B. Hariharan. "Pointflow: 3d point cloud generation with continuous normalizing flows". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 4541–4550.

[185] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. "Generative adversarial text to image synthesis". In: *International conference on machine learning*. PMLR. 2016, pp. 1060–1069.

[186] J. Zhu, Z. Wang, J. Chen, Y.-P. P. Chen, and Y.-G. Jiang. "Balanced Contrastive Learning for Long-Tailed Visual Recognition". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 6908–6917.

[187] J. Cui, Z. Zhong, S. Liu, B. Yu, and J. Jia. "Parametric contrastive learning". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 715–724.