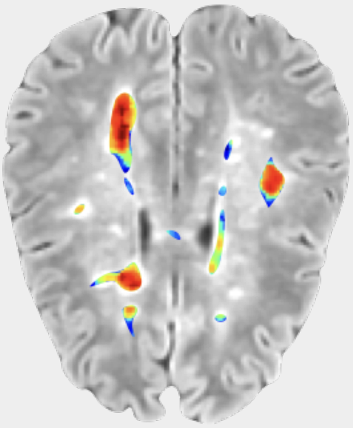
Computer Aided Medical Procedures
Prof. Dr. Nassir Navab



Dissertation

Anomaly Detection in Brain MRI: From Supervised to Unsupervised Deep Learning

Christoph Baur

Fakultät für Informatik
Technische Universität München

Technische Universität München
Fakultät für Informatik
Lehrstuhl für Informatikanwendungen in der Medizin

Anomaly Detection in Brain MRI: From Supervised to Unsupervised Deep Learning

Christoph Baur

Vollständiger Abdruck der von der Fakultät für Informatik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzende(r):	Prof. Dr. Julien Gagneur
Prüfer der Dissertation:	1. Prof. Dr. Nassir Navab
	2. Dr. Ben Glocker

Die Dissertation wurde am 28.10.2020 bei der Technischen Universität München eingereicht und durch die Fakultät für Informatik am 17.03.2021 angenommen.

Christoph Baur
Anomaly Detection in Brain MRI: From Supervised to Unsupervised Deep Learning
Dissertation, Version 1.0

Technische Universität München
Fakultät für Informatik
Lehrstuhl für Informatikanwendungen in der Medizin
Boltzmannstraße 3
85748 Garching bei München

Abstract

Anatomical imaging of the human brain and central nervous system is a fundamental component of todays diagnosis and therapy of various neurological diseases. To relate symptoms to neurological causes, radiologists analyze the high resolution visualizations provided by modalities such as Magnetic Resonance Imaging (MRI) and try to identify abnormal structures. This manual process is not only cumbersome, time-consuming and costly, but also prone to human error: According to numerous studies, in up to 5-10% of images pathologies remain unnoticed. Lately, breakthroughs in the automatic, computer-assisted analysis of brain MRI achieved through advances in the field of machine learning—so-called supervised Deep Learning methods—have shown performances on par with or even outperforming that of human experts. However, these methods come at a cost: Such deep artificial neural networks need to be trained from vast amounts of carefully annotated examples of benign and diseased cases, for which manual curation and thus precious expert human resources are required. Moreover, these methods do not provide guarantees of being capable of identifying pathologies not present in the training data. The research presented in this thesis focuses on overcoming these burdens and highlights a path from the paradigm of supervised Deep Learning towards annotation-free, unsupervised methods. This includes i) semi-supervised concepts which can leverage both annotated and unlabeled data to improve generalization of Deep Learning-based approaches, and ii) unsupervised anomaly detection frameworks which do not require manual markings of pathologies at all. The latter contributions employ deep representation learning, generative modeling and image-to-image translation techniques to build a model of normal anatomy, which allows to identify anomalies in brain MRI as distributional outliers. This way, opposed to supervised methods, the resulting models are not pathology-specific. A major emphasis is further put on modeling healthy brain distributions at high resolution to be able to detect and delineate particularly small brain lesions. Ultimately, it is shown that the fusion of aforementioned supervised and unsupervised techniques yields an effective self-teaching framework for brain lesion segmentation. This framework also outlines a potential, inexpensive way of integrating Deep-Learning-based anomaly detection into everyday clinical routine. Although validated on brain MRI, the concepts can certainly be translated to other imaging modalities and parts of the anatomy (e.g. CT, X-Ray), opening up great opportunities for integration of Deep Learning into radiology.

Zusammenfassung

Die anatomische Bildgebung des menschlichen Gehirns sowie des zentralen Nervensystems ist ein fundamentaler Bestandteil der gegenwärtigen Diagnosestellung und Therapie zahlreicher neurologischer Pathologien. Um Symptome und neurologische Ursachen zu korrelieren, analysieren Radiologen das durch bildgebende Verfahren wie MRT bereitgestellte, hochauflösende Bildmaterial und versuchen darin abnormale Strukturen zu identifizieren. Dieser manuelle Prozess ist nicht nur mühsam, zeitaufwändig und kostspielig, sondern auch fehlerbehaftet: Mehrere Studien zeigen auf, dass in 5-10% der Bilder Pathologien übersehen werden. Aktuelle Durchbrüche in der automatischen, Computer-gestützten Analyse von Gehirn-MRT mittels Maschinellem Lernen—so genannte Supervised Deep Learning Verfahren—ermöglichen eine Performanz ähnlich oder besser der von menschlichen Experten. Allerdings gehen solche Deep-Learning-Ansätze auch mit Nachteilen einher: Das Training von tiefen, künstlichen neuronalen Netzen erfordert große Mengen an sorgfältig annotierten Beispieldaten von gutartigen und pathologischen Fällen, die manuell gesammelt werden müssen und deshalb kostbare Ressourcen von Experten erfordern. Darüber hinaus garantieren diese Ansätze nicht, beliebige Anomalien identifizieren zu können, die in dieser Art nicht in den Trainingsdaten vorkommen. Die im Rahmen dieser Dissertation präsentierten Forschungsergebnisse versuchen diese Probleme zu überwinden und zeigen einen Weg vom Paradigma des Supervised Deep Learning zu unsupervidierten Methoden, die keine annotierten Daten benötigen. Dies umfasst insbesondere i) Semi-supervised Deep Learning Ansätze, die sowohl von annotierten als auch nicht-annotierten Daten lernen und die Generalisierung von Deep-Learning-basierten Methoden verbessern können, und ii) so genannte Unsupervised Anomaly Detection Verfahren, die keine manuellen Annotationen von Pathologien benötigen. Die zugrundeliegenden Techniken umfassen Deep Representation Learning, Generative Modeling und Image-to-Image Translation Techniken um ein Modell gesunder, normaler Anatomie zu lernen, was im Nachgang die Identifikation von Anomalien als Verteilungsausreißer erlaubt. Im Gegensatz zu supervidierten Methoden sind die gelernten Modelle nicht Pathologie-spezifisch. Die vorgestellte Forschung legt darüber hinaus einen besonderen Schwerpunkt auf das Modellieren der Verteilung des gesunden menschlichen Gehirns mit hochauflösenden Daten, um das Detektieren und Segmentieren von besonders kleinen Gehirnläsionen zu ermöglichen. Zuletzt wird aufgezeigt, dass die Fusion von supervidierten und unsupervidierten Techniken ein effektives, selbst-lernendes Framework für die Segmentierung von Gehirnläsionen ergibt. Dieses Framework zeichnet auch potentielle Wege für eine kosteneffektive Integration von Deep Learning-basierten Anomalie-Erkennungsverfahren in den klinischen Alltag auf. Die vorgestellten Konzepte werden zwar am Beispiel von Gehirn-MRT-Daten validiert, können aber potentiell auch auf andere bildgebende Verfahren (z.B. CT und Röntgen) und Teile der Anatomie angewandt werden. Dies bedeutet große Chancen für die Integration von Deep Learning in das Umfeld der Radiologie.

Acknowledgments

To start, I would like to thank my PhD supervisor, Prof. Dr. Nassir Navab, for giving me the opportunity to pursue the degree of a PhD. I am not only thankful for the ability to work on this exciting project, but for bringing me and the company of R&S together and thus shaping my career so substantially. I am very thankful for the chance to get to know so many brilliant people all over the world, both at the many conferences and gatherings of the medical image analysis community and directly at the Chair for Computer Aided Medical Procedures (CAMP). My two years as a scientfic employee at the chair were demanding, but also exciting and much fun. This is especially because of my peers at the chair, many of whome became close friends. Anees, Mai and Hendrik and Johanna, thanks for making the time at the chair so worthwhile. Mai & Anees have been the best office mates once could possibly imagine. Markus, thanks for all the years of great friendship and the exciting journey through higher education which we fought together from Abitur to PhD. I'd also like to thank my mentor, Dr. Shadi Albarqouni, for his endless support, the countless fruitful, scientific discussions, his precious feedback and trust in my capabilities. Throughout the years, we have maintained a great, prosperous scientfic relationship. I highly appreciate the ability to freely choose this research topic and follow my own instincts. I would also like to thank Dr. Benedikt Wiestler from the Neuroradiology department of Klinikum Rechts der Isar for the close scientific collaboration, which resulted in numerous scientific contributions. He did not only generously provide data for our joint research endeavors, but committed a lot to multiple projects with ideas and his admirable writing skills. Thanks to my students, especially Robert and Stefan, for fighting in the projects we did together and investing so much of their valuable time.

The company of Rohde & Schwarz also deserves an equally strong, honorable mention. In particular, I would like to thank Dr. Athanasios Karamalis and Christian Evers for bringing me on board to their brilliant team and for generously funding my research. I am incredibly thankful for them offering me a position inside the team and for heavily supporting my personal growth within the company. I would have never imagined my career to evolve so rapidly.

Last but not least, I want to thank all of the great human beings which mean so much to me and never stopped supporting me with their care and love, even though the physical presence I could offer to them was so limited. Thanks to Lisa, who gave me her love and yet all the freedom I needed to be able to finish this PhD. I also want to thank my parents and my brother, my grand-parents and of cause my close friends who also have always been there for me when I needed them. They had an ear for all my problems, listened to the complaints when I struggled, and shared my joy when I had success.

Contents

List of Abbreviations

AAE Adversarial Autoencoder. 23, 31

AD Anomaly Detection. xix, 55, 59–61

AE Autoencoder. 6, 19–21, 30, 31, 34, 36, 38, 60, 71

ANN Artificial Neural Network. 9–11, 19, 20, 23

CAD Computer-Aided Diagnosis. 59

cGAN Conditional Generative Adversarial Network. 24, 25

CNN Convolutional Neural Network. 11–15, 71

CRF Conditional Random Field. 14

CSF Cerebro-Spinal Fluid. 4

CT Computed Tomography. 2, 6, 13, 14, 59

DCGAN Deep Convolutional Generative Adversarial Network. 24, 26, 27

DL Deep Learning. xix, 1, 2, 5–7, 10–15, 19, 23, 31, 55, 59–61

DNN Dense Neural Network. 9, 21, 24

ELBO Evidence Lower-Bound. 22

EM Expectation Maximization. 6

FCN Fully Convolutional Network. 12–15, 71

FLAIR Fluid-Attenuated Inversion Recovery. 4, 6

FP False Positive. 60, 61

GAN Generative Adversarial Network. 19, 23–27, 30, 31, 38, 60, 71

GCN Graph Convolutional Network. 59

GMVAE Gaussian Mixture Variational Autoencoder. 23, 31

KNN K-Nearest Neighbor. 6

MC Monte-Carlo. 31, 34

MLP Multilayer Perceptron. 5, 7–9

MR Magnetic Resonance. 1, 3, 5, 6, 13, 14, 23, 27, 31, 34, 60, 71

MRF Markov Random Field. 6

MRI Magnetic Resonance Imaging. v, xix, 1–6, 13, 14, 27, 30, 31, 34, 38, 55, 59, 61, 71

MS Multiple Sclerosis. 1, 5, 6, 14, 15, 31, 36, 61, 71

OCT Optical Coherence Tomography. 6, 26, 30

OP Operating Point. 60

PET Positron Emission Tomography. 2

RF Radiofrequency. 2, 3

SGD Stochastic Gradient Descent. 10, 11

SOM Self-Organizing Map. 6

SPECT Single Photon Emission Computed Tomography. 2

SVM Support Vector Machine. 6, 30

T1w T1-weighted. 4

T2w T2-weighted. 4

UAD Unsupervised Anomaly Detection. xii, xix, 19, 26, 30, 31, 55, 59–61

UAS Unsupervised Anomaly Segmentation. 19, 30, 31, 36, 38, 55, 71

VAE Variational Autoencoder. 19, 21–23, 25, 26, 30, 31, 36, 38, 60, 71

WMH White-Matter Hyperintensity. 5, 6, 31

WML White-Matter Lesion. 5, 6, 14, 34, 55, 56

List of Authored and Co-authored Publications

2020

[12] **Christoph Baur**, Robert Graf, Benedikt Wiestler, Shadi Albarqouni, and Nassir Navab. "SteGANomaly: Inhibiting CycleGAN Steganography for Unsupervised Anomaly Detection in Brain MRI." *In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 718-727. Springer, Cham, 2020.*

[17] **Christoph Baur**, Benedikt Wiestler, Shadi Albarqouni, and Nassir Navab. "Scale-Space Autoencoders for Unsupervised Anomaly Segmentation in Brain MRI." *In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 552-561. Springer, Cham, 2020.*

[11] **Christoph Baur**, Stefan Denner, Benedikt Wiestler, Shadi Albarqouni, and Nassir Navab. "Autoencoders for Unsupervised Anomaly Segmentation in Brain MR Images: A Comparative Study." *arXiv preprint arXiv:2004.03271 (2020).*

[14] **Christoph Baur**, Benedikt Wiestler, Shadi Albarqouni, and Nassir Navab. "Bayesian Skip-Autoencoders for Unsupervised Hyperintense Anomaly Detection in High Resolution Brain Mri." *In 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), pp. 1905-1909. IEEE, 2020.*

[53] Salome Kazeminia, **Christoph Baur**, Arjan Kuijper, Bram van Ginneken, Nassir Navab, Shadi Albarqouni, and Anirban Mukhopadhyay. "GANs for medical image analysis." *Artificial Intelligence in Medicine (2020): 101938..*

2019

[16] **Christoph Baur**, Benedikt Wiestler, Shadi Albarqouni, and Nassir Navab. "Fusing unsupervised and supervised deep learning for white matter lesion segmentation." *In International Conference on Medical Imaging with Deep Learning, pp. 63-72. 2019.*

[23] Mai Bui, Felix Bourier, **Christoph Baur**, Fausto Milletari, Nassir Navab, and Stefanie Demirci. "Robust navigation support in lowest dose image setting." *International journal of computer assisted radiology and surgery 14, no. 2 (2019): 291-300.*

[22] Mai Bui, **Christoph Baur**, Nassir Navab, Slobodan Ilic, and Shadi Albarqouni. "Adversarial Networks for Camera Pose Regression and Refinement." *In Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 0-0. 2019.*

[92] M. Tarek Shaban, **Christoph Baur**, Nassir Navab, and Shadi Albarqouni. "Staingan: Stain style transfer for digital histological images." *In 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pp. 953-956. IEEE, 2019.*

2018

[15] **Christoph Baur**, Benedikt Wiestler, Shadi Albarqouni, and Nassir Navab. "Deep autoencoding models for unsupervised anomaly segmentation in brain MR images." *In International MICCAI Brainlesion Workshop, pp. 161-169. Springer, Cham, 2018.*

[9] **Christoph Baur**, Shadi Albarqouni, and Nassir Navab. "MelanoGANs: high resolution skin lesion synthesis with GANs." *arXiv preprint arXiv:1804.04338 (2018).*

[8] **Christoph Baur**, Shadi Albarqouni, and Nassir Navab. "Generating highly realistic images of skin lesions with GANs." *In OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis, pp. 260-267. Springer, Cham, 2018.*

2017

[10] **Christoph Baur**, Shadi Albarqouni, and Nassir Navab. "Semi-supervised deep learning for fully convolutional networks." *In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 311-319. Springer, Cham, 2017.*

2016

[1] Shadi Albarqouni, **Christoph Baur**, Felix Achilles, Vasileios Belagiannis, Stefanie Demirci, and Nassir Navab. "Aggnet: deep learning from crowds for mitosis detection in breast cancer histology images." *IEEE transactions on medical imaging 35, no. 5 (2016): 1313-1321.*

[7] **Christoph Baur**, Shadi Albarqouni, Stefanie Demirci, Nassir Navab, and Pascal Fallavollita. "CathNets: detection and single-view depth prediction of catheter electrodes." *In International Conference on Medical Imaging and Augmented Reality, pp. 38-49. Springer, Cham, 2016.*

2015

[13] **Christoph Baur**, Fausto Milletari, Vasileios Belagiannis, Nassir Navab, and Pascal Fallavollita. "Automatic 3D reconstruction of electrophysiology catheters from two-view monoplane C-arm image sequences." *International journal of computer assisted radiology and surgery 11, no. 7 (2016): 1319-1328.*

Outline of the thesis

The introductory part of this thesis provides the motivation behind the presented research as well as a background in medical paradigms of MRI and the different pathologies of interest. A brief chronological review of Anomaly Detection (AD) in the field of brain imaging gently introduces the reader into the paradigm.

The second part of this thesis focuses on Supervised Deep Learning in the context of AD in brain MRI. To stay self-contained as much as possible, this chapter also provides some historical and theoretical background on Deep Learning (DL) before selected works and a contribution to this field are presented.

The third and biggest part of this work concentrates on AD methods in the realm of Unsupervised Deep Learning. Deep unsupervised representation learning and generative modeling techniques are the very foundation of this work and thus first introduced. Afterwards, different conceptual categories for UAD are presented alongside various contributions.

Last, a method which combines the best of both supervised and unsupervised Deep Learning methods into an effective, hybrid pipeline is displayed. In light of this contribution, discussions on clinical utility and open challenges are triggered and perspectives for future work are given.

Introduction

1

Our most exciting discoveries come from studying anomalies. The once-in-1000 occurrence is worth getting detail on.

— **Michael J. Silverstein**

1.1 Motivation

The acquisition and analysis of anatomical images of the human brain and central nervous system is an essential part of every-day life of Neuro-radiologists. Therein, MRI is a commonly used, non-invasive and radiation-free key technology that provides the clinician with a high-fidelity visualization of the anatomy of interest. A vital part of diagnosis, progression monitoring and treatment of various neurological diseases is the manual analysis and identification of abnormalities in the volumetric data Magnetic Resonance (MR) imaging delivers. Often times, a delineation of brain lesions is also necessary to aid in the quantification of progression- or stage-predictive biomarkers [48], and manual segmentation is still considered the gold-standard. Unfortunately, manual processing by domain experts is cumbersome, time-consuming and costly. It is also an error-prone process: According to [21], in up to 5-10% of cases pathologies remain undiscovered. Further, considerable inter-observer-disagreement has been reported [19, 24].

The medical image analysis community has made countless efforts towards automating these steps, striving for human level precision at economically acceptable computational complexity. Lately, breakthroughs could be made thanks to major advances in the field of machine learning. So-called supervised DL methods have shown outstanding performance in the detection and segmentation of pathologies such as Multiple Sclerosis (MS) lesions, brain tumors and ischemias. In fact, scientific results indicate capabilities on par with human experts [65].

Supervised Deep Learning leverages vast amounts of annotated training data—usually thousands of curated examples—in order to solve complex, diverse image analysis tasks. Once trained, the resulting models are experts in dealing with the tasks they have been optimized for. Instead of requiring extensive domain knowledge to hand-craft features descriptive enough to solve specific problems, supervised DL learns to extract relevant features directly from data in a hierarchical, non-linear fashion. As such, the amount, content and quality of annotated training data are crucial factors. While the resulting supervised approaches perform well in the classification and segmentation of various diseases, they are primarily pathology-specific and do not provide guarantees to generalize indefinitely beyond the cases they have been trained for. Overall, annotations define the boundaries of optimization objectives such as disease classification or segmentation. However, in the medical field, annotations are usually the result of a tedious manual process, in which domain experts sift through examples and assign markings on different levels, ranging from per-sample labels up to a granularity of single voxels. In summary, supervised DL involves two challenges: i) the costly and time-consuming nature of manual expert labelling leads to a scarcity of annotated data; ii) further, training data seldomly covers all possible pathological phenotypes.

1.2 Challenges & Contributions

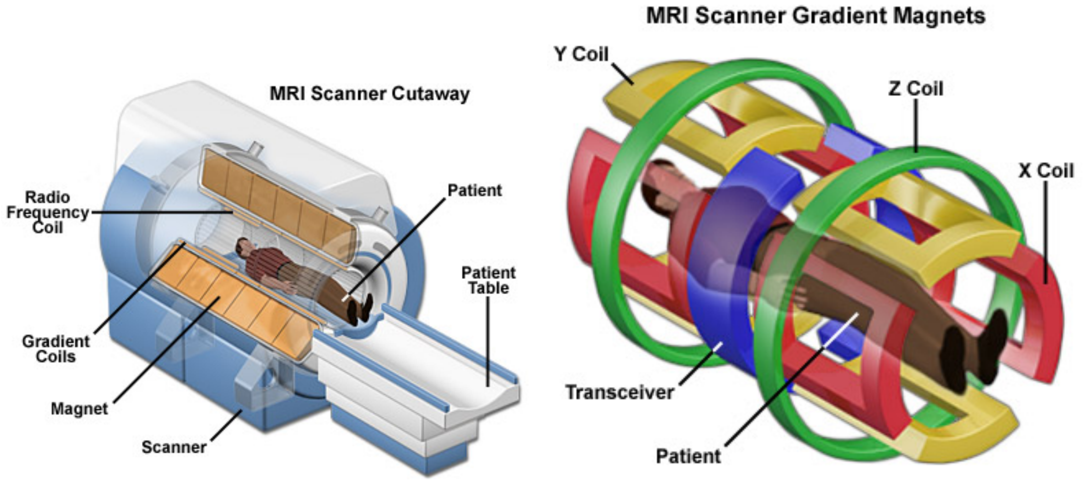
The research presented in this thesis focuses on overcoming these burdens and highlights a path from the paradigm of supervised DL towards annotation-free, unsupervised DL methods. This includes i) semi-supervised concepts which can leverage both annotated and unlabeled data to improve generalization of DL-based approaches, and ii) unsupervised anomaly detection frameworks which do not require manual markings of pathologies at all. The latter contributions employ deep representation learning, generative modeling and image-to-image translation techniques to build a model of normal anatomy, which allows to identify anomalies in brain MRI as distributional outliers. Identification of abnormalities thereby ranges from a per-sample basis to pixel-precise segmentations, and the resulting models are not pathology-specific. A major emphasis in this work is further put on modeling healthy brain distributions at high resolution to be able to detect and delineate particularly small brain lesions with great precision. Last, it is shown that the fusion of aforementioned supervised and unsupervised techniques can serve as an effective self-supervised framework for brain anomaly segmentation. This framework also outlines a potential, inexpensive way of integrating DL-based anomaly detection into everyday clinical routine. Although validated on brain MRI, the concepts can certainly be translated to other imaging modalities and parts of the anatomy (e.g. CT, X-Ray), opening up great opportunities for bringing Deep Learning into radiology.

1.3 Brain Imaging with Magnetic Resonance

1.3.1 Physical Basics

MRI is a non-invasive imaging modality and as opposed to X-ray, Computed Tomography (CT), Positron Emission Tomography (PET) or Single Photon Emission Computed Tomography (SPECT) imaging does not expose the patient to ionizing radiation. The imaging relies on the measurement of weak magnetic fields and their inductive properties emitted by hydrogen atoms when subjected to specific, sudden changes of a strong, surrounding magnetic field. As the human body has a water content of approx. 70% with varying hydrogen compositions among tissue-types, different signal strength of the emitted weak magnetic fields can be measured, providing great contrast of soft-tissue.

MRI Building Blocks—The MRI scanner (see Fig. 1.1) commonly consists of a primary magnet with a strong magnetic field. Within the bore of the primary magnet, so-called gradient magnet coils are located, which are aligned orthogonally to each other and alter the primary magnetic field in every direction. Radiofrequency (RF) coils constitute the third



(a) Schematic of an MR scanner

(b) Illustration of gradient and radiofrequency coils in an MR scanner

Fig. 1.1 Graphic illustrations of the building blocks of an MR scanner (Source: [32])

essential building block, serving both as RF pulse transmitters and signal receivers during the image acquisition. Last, a computer system is required for turning the signals received from the RF receivers into an actual image.

Hydrogen—At the heart of MR imaging is the hydrogen atom, whose nucleus is solely composed of a single, positively charged proton. As a spinning, charged particle, it possesses a magnetic field or a so-called magnetic moment, equiping the proton with a direction. Along this axis, the proton spins like a spinning top at a certain precession rate, also known as the Larmor Frequency. In a normal environment, protons are oriented randomly, such that there is no overall magnetic field produced by a mass of hydrogen atoms, and they do not precess together, i.e. they are out-of-phase. However, the precession rate changes proportionally to the strength of the surrounding magnetic field.

Principle—The strong primary magnetic field $\mathbf{B}_0$ forces the protons of hydrogen atoms to align either parallel (low-energy state) or anti-parallel (high-energy state) to the fields direction ("longitudinal magnetization"). Most of protons will align parallel, such that the net magnetic field (the sum of all magnetic field directions) of the parallel and anti-parallel protons is in the direction of the primary magnetic field. The aforementioned gradient coils create secondary fields which give MRI the capability to image directionally along the x, y and z axis with one gradient coil for every axis. A spatial encoding is ultimately facilitated through resulting local changes of the primary magnetic field and hence local precession rates of protons. In order to measure signals, RF pulses are applied at specific precession frequencies, which create temporary disturbances in the alignments of protons. Some low-energy protons flip to a high-enery state, decreasing longitudinal magnetization, and start to precess in-phase, turning the net magnetic vector orthogonal to the longitudinal magnetic field ("transverse magnetisation"). After the pulse has been applied, protons spiral back to their initial state inside the primary magnetic field. The simultaneously changing magnetic moment of the net magnetic vector results in a free induction decay, inducing an electrical signal measured by the RF coils. The time it takes for protons to reach their inertial equilibrium state is known as *relaxation* and essentially can be measured along the longitudinal axis (“T1 relaxation”) and in the transverse axis (“T2 relaxation”). As relaxation time is contingent on tissue, RF coils receive different signals for different parts of the anatomy. The received signals are digitized and collected in the so-called k-space representation, from which actual images are recovered with the help of the inverse Fourier transform.

Since MR imaging primarily relies on hydrogen nuclei and their spins, the imaging is well suited for visualizing soft-tissue at high contrast and resolution. However, MR imaging prohibits the exposure of any metallic objects to the strong magnetic field, e.g. limiting its applicability to patients without metallic implants. Long acquisition times and the common usage of sterile, metallic medical instruments further render it inapplicable for live imaging during interventions.

1.3.2 MRI Sequences

MRI sequences leverage different proton spin relaxations to provide better contrast between specific tissue types. Sequences utilized in clinical routine commonly comprise T1-weighted (T1w), T2-weighted (T2w), and Fluid-Attenuated Inversion Recovery (FLAIR).

T1w—T1 relaxation describes the re-alignment of proton spins to the primary magnetic field B_0, which varies for different tissue. Fat tissue has very fast relaxation, i.e. realigns very fast with B_0, and appears bright (“hyper-intense”) in T1-weighted images. Opposed to that, water realigns much slower to the longitudinal magnetization upon casting an RF pulse and has low intensity (“hypo-intense”) in the resulting images. Since the white matter of the human brain has a lot of myelin content, made up of complex fat molecules, it shows overall high intensity, whereas Cerebro-Spinal Fluid (CSF) is dark. T1w imaging is suitable for anatomical imaging and allows to visualize vascular changes as well as disruptions of the blood-brain-barrier. It also provides great contrast for paramagnetic contrast agents such as Gadolinium.

T2w—T2 relaxation captures tissue-dependent differences in the spin decay of protons with aligned precession (in-phase) while returning to their intertial state. In comparison to T1, T2 relaxation provides inverted signal intensity for fat tissue and water, i.e. fat appears to be dark and water has high signal intensity. A very useful property of the resulting contrast is the hypo-intensity of fluids flowing with high velocity (such as blood), which allows to visualize blood vessels. T2w imaging is well suited for making a large variety of brain lesions visible. However, if lesions are close to the CSF, it is hard to tell them apart.

FLAIR—The FLAIR sequence constitutes one particular example of so-called inversion recovery sequences in which multiple, specific RF pulses are employed as to annihilate the signal for selected tissues. FLAIR images are generally similar to T2, but the signal of free-flowing fluids is suppressed such that CSF becomes hypo- instead of hyper-intense. This allows to clearly delineate lesions near the ventricles of the brain. In addition, FLAIR generally provides improved differentiation between white- and gray matter.

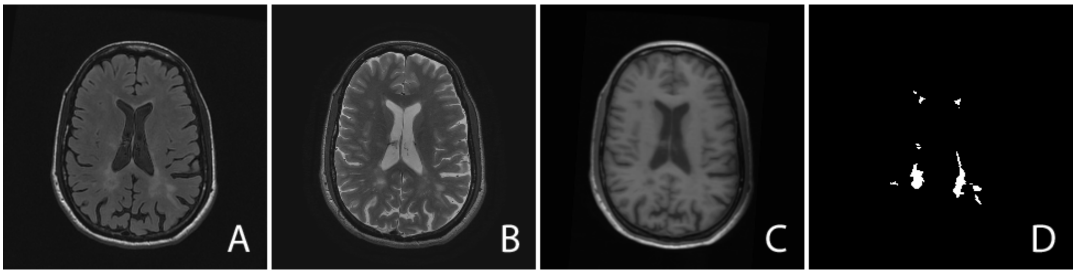


Fig. 1.2 The same axial MRI slice containing MS lesions in three different protocols. A: FLAIR; B: T2; C: T1; D: Ground-truth segmentation of the lesions (slices taken from a sample of the MSSEG 2008 dataset).

1.3.3 Applications in Brain Imaging

MRI is commonly employed in the diagnosis and progression monitoring of brain tumors and cysts, various neurological diseases such as MS, Alzheimers or epilepsy, inflammations and infections, hemorrhaeges and strokes, vascular diseases, morphological malformations, developmental anomalies and traumatic brain injuries. As such, deviations from the norm of human brains can be manifold, ranging from characteristic intensity abberations to significant structural- and morphological changes. So-called White-Matter Lesions (WMLs) constitute prominent intensity-characteristic biomarkers for a plethora of pathological, physiological processes such as inflammation, demyelination, axonal loss or edema [39]. In turn, these are known to be causes of diseases such as MS, Alzheimers, dementia or epilepsy. White-Matter Lesions vary in shape, size and location, which is in many cases contingent on the pathology. In FLAIR scans of subjects suffering from such pathologies, White-Matter Lesions appear as particularly strong, high magnitude signals and are hence often referred to as White-Matter Hyperintensitys (WMHs) (see Fig. 1.2 for a comparison of different MRI contrasts from a subject with MS). Acute ischemias and bleedings as the result of strokes, ruptured aneurysms or traumatic brain injuries are commonly diagnosed with diffusion-weighted MR imaging and are not contingent on specific anatomical regions. Brain tumors and their different compartments comprise different tissue-types with varying properties (tumor-induced edema near the tumor itself, active and necrotic tumor tissue) and may induce morphological deformations to the brain.

1.4 A Brief History of Anomaly Detection in Brain Imaging

Computer-assisted brain image analysis has seen countless efforts towards anomaly detection over the last decades, ranging from methods which i) do not rely on any prior knowledge over ii) approaches which model only normality iii) to techniques which leverage knowledge about both normal and abnormal cases [94]. Before the rise of DL, anomaly detection in medical image analysis employed parametric [38, 58, 97] and non-parametric statistical modeling [72], pattern recognition, content-based retrieval [27], clustering and outlier detection techniques as well as machine-learning techniques established at that time. Some early, data-driven attempts combined traditional image processing algorithms with Multilayer Perceptrons (MLPs) and Self-Organizing Maps (SOMs). A variety of such “traditional” approaches for anomaly detection in medical imaging is surveyed in [94] and evaluated on a benchmark CT dataset.

In the specific context of WML detection and segmentation in brain MRI, works prior to DL used parametric statistical models in the form of Markov Random Fields (MRFs) [97], adaptive thresholding and 3D connected component analysis [44], probabilistic K-Nearest Neighbor (KNN) techniques [3] or fuzzy clustering techniques combined with topological constraints [93]. Machine-learning-based works revolved around encoding representations of normal brain MR patches with Dictionary Learning and Sparse-Coding techniques for detecting MS lesions [99]. BIANCA [41] (Brain Intensity AbNormality Clustering Algorithm) is another recent supervised, k-Nearest-Neighbor-based method for detecting and delineating abnormal White-Matter Hypterintensities in arbitrary MRI modalities. MSMetrix [46] uses spatial, anatomical- and intensity priors in an Expectation Maximization (EM)-based WML-detection algorithm for 3D FLAIR. LST [88, 89] embeds MS lesion segmentation in a bayesian MRF framework.

Compared to WML, brain tumor segmentation has been a more active field. Presumably due to a lack of a sufficiently large gold-standard, various early approaches relied on modeling only normal anatomy. For this purpose, pioneering work employed 3D hierarchical deformable registration to an ATLAS [27]. Similarly, [72] used a registered brain atlas as a model for normality together with robust estimation techniques to segment brain tumors. Menze et al. developed a generative model on probabilistic atlases [66]; [38] fused diagonalized nearest neighbor pattern recognition, MRF and iterated bayesian classification exclusively trained on healthy tissue.

With the rise of DL, manual mathematical modeling and feature-crafting has been pushed to the backseat and a great variety of primarily data-driven approaches emerged. As in many other fields, Supervised DL has quickly set new standards in the detection, classification and segmentation of diseases such as MS, Alzheimers, tumors and ischemias in brain MRI [51, 96, 105]. The success of those mostly pathology-specific methods has been highlighted in many different challenges such as BraTS [65] or the 2015 longitudinal MS lesion segmentation challenge [24]. Nonetheless, the need for unsupervised, less annotation-intensive solutions has been quickly identified. Early work in this realm used the unique capability of Autoencoders (AEs) to learn non-linear projections of normal anatomy onto a lower-dimensional manifold and one-class Support Vector Machines (SVMs) for clustering-based detection of epilepsy lesions [37] and retinal abnormalities in Optical Coherence Tomography (OCT) data [91]. More recent works rely on AEs and deep generative modeling for pixel-precise localization and segmentation of abnormal structures directly in the input space. The contributions presented in this work also fall into this category.

2 Supervised- & Semi-supervised Deep Learning for Anomaly Detection in Brain MRI

2.1 Definition of Brain Anomaly Detection in a Supervised Context

In their survey of traditional anomaly detection papers for medical image analysis, Taboada-Crispi et al.[94] differentiate between three different types of anomaly detection, based on degress of prior knowledge: a) methods completely devoid of prior knowledge, b) approaches which know alone about normality and c) techniques which utilize knowledge about both normal and abnormal cases. Supervised DL falls within the third category, as it generally requires both healthy and anomalous samples to be able to tell them apart. Based on the type and degree of provided labels, different objectives can be formulated. Each sample may be assigned a specific class label from the set of classes $C = \{\text{normal}, \text{anomalous}\}$, framing anomaly detection as a binary classification task. Alternatively, if the (rough) location of anomalies within all samples is provided, anomaly detection may be defined as a localization task to regress the coordinates of anomalies. In the most extreme case, every pixel of each sample is labeled as either healthy or anomalous, such that anomaly detection may be performed through a pixel-wise segmentation objective.

Hence, in addition to the degrees of prior knowledge involved, three different sub-types of anomaly detection are distinguished in this work (not exclusive to supervised DL): 1) identification of anomalous samples, 2) anomaly localization and 3) anomaly segmentation, all of which can be formulated as supervised DL tasks. Cases 2) and 3) may be interpreted as more specialized variants of 1) and are primarily considered in this work.

2.2 Supervised Deep Learning

2.2.1 The Multilayer Perceptron

The very foundation of all DL is the so-called Perceptron [79] and its more complex, non-linear variant, the MLP. Dating back to the mid of the 20th century, the Perceptron pursued the formulation of a basic mathematical analogue to the biological neurons found in human and animal nervous systems. The ordinary neuron known from biology (see Fig. 2.1) gathers chemical signals released by other incoming neurons at its dendrites, transforms them into electrical potentials and accumulates these. Once the accumulation exceeds a so-called

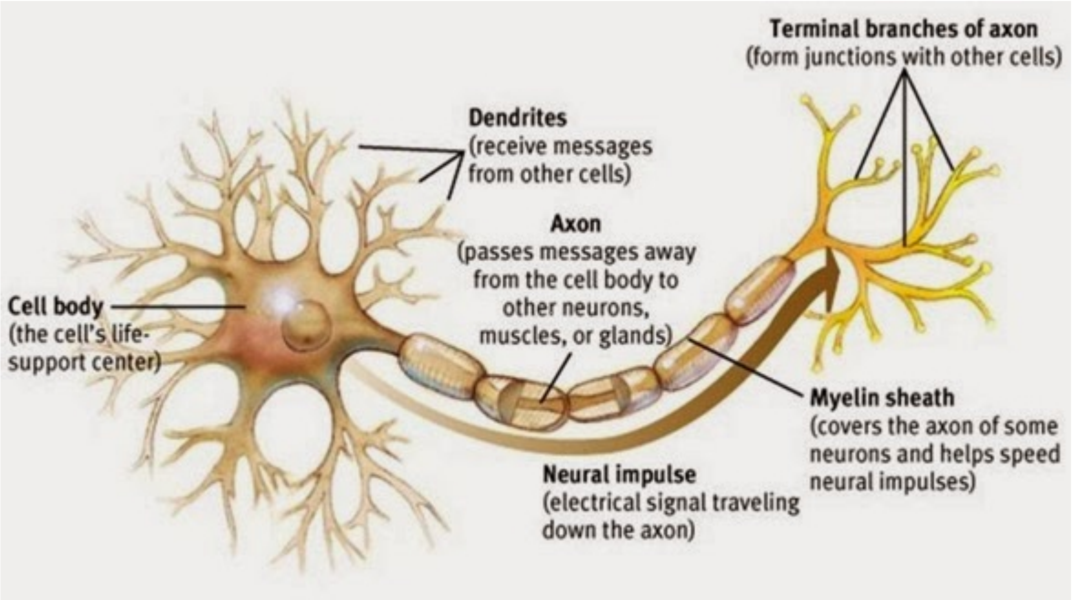


Fig. 2.1 Illustration of the biological motor-neuron (Source. https://hspersunite.org.au/neurons-more-complex-than-thought/, August 23rd 2020)

activation potential, the neuron fires an electrical signal along its axon and propagates the signal to connected neurons by releasing chemicals into the synaptic gap. Inspired by this, the Perceptron (Fig. 2.2) models such neural behavior with a simple function $f(\mathbf{x})$, in which neural input information is represented as a vector $\mathbf{x} \in \mathbb{R}^N$. Another vector $\mathbf{w} \in \mathbb{R}^N$ and a so-called bias term b analogously capture the neuron-specific eletrical potentials. The resulting mathematical function $f(\mathbf{x})$ mimics neural behavior with a simple linear combination of $\mathbf{w}$, b and $\mathbf{x}$:

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

The weights $\mathbf{w}$ and bias b constitute trainable parameters which can be *learned* to make $f(\cdot)$ approximate specific linear functions or to solve binary classification problems which are linearly separable. Traditional ways to train the parameters involve solving a least-squares optimization problem based on a set of training data $\mathbf{X}$ with corresponding labels $\mathbf{Y}$.

To extend the capabilities of the linear Perceptron to approximate complex, non-linear functions, the so-called MLP (Fig. 2.3) has been introduced shortly thereafter. The MLP stacks multiple, but at least two layers of combinations of Perceptrons and non-linear, differentiable activation functions $\sigma(\cdot)$ to form a composite function of L layers:

$$f(\mathbf{x}) = \sigma(\mathbf{w}_L^T(...\sigma(\mathbf{w}_1^T\sigma(\mathbf{w}_0^T\mathbf{x} + b_0) + b_1) + b_L) \tag{2.2}$$

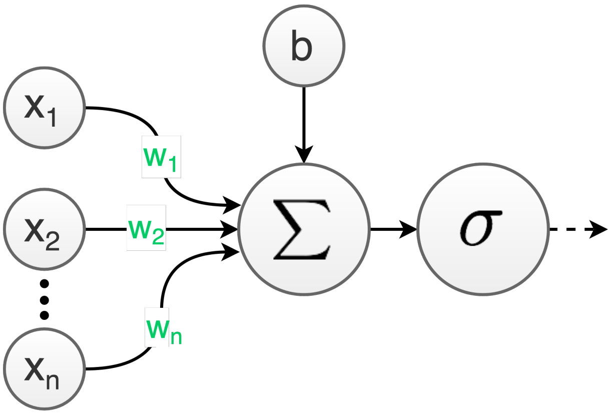


Fig. 2.2 Graphical illustration of the Perceptron.

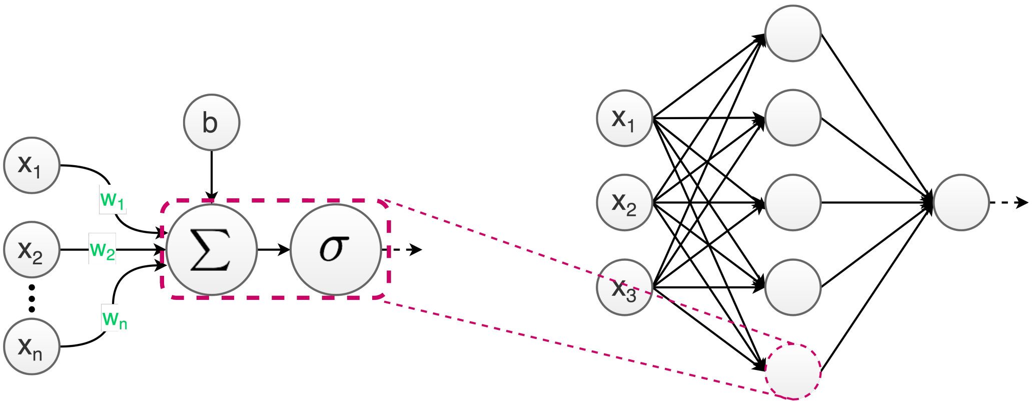


Fig. 2.3 Graphical illustration of the Multi-layer Perceptron

Layers 0 to $L-1$ are commonly denoted as *hidden layers*, and layer L represents the output layer of the MLP. Popular activation functions comprise the sigmoid function (2.3), hyperbolic tangent (2.4) and the piece-wise linear rectified linear unit (2.5):

$$\sigma(\mathbf{x}) = \frac{1}{1 + \mathrm{e} - \mathbf{x}} \tag{2.3}$$

$$\sigma(\mathbf{x}) = tanh(\mathbf{x}) \tag{2.4}$$

$$\sigma(\mathbf{x}) = max(0, f(\mathbf{x})) \tag{2.5}$$

Today, the Perceptron and MLP are a part of the much broader family of Artificial Neural Networks (ANNs). Therein, the topology of the MLP is also known as Dense Neural Network (DNN) because of the dense, any-to-any connectivity among neurons of neighboring layers. Although those concepts have been around since the mid of the 20th century, substantially successful applications could only be reported recently. This success is grounded i) on the contemporary efficacy, efficiency and large-scale availability of computational resources, ii) the availability of a wide range of composite, specialized building blocks as well as iii) the discovery of an efficient, tractable training algorithm: backpropagation [82] with Stochastic Gradient Descent (SGD).

2.2.2 Backpropagation

The parameters of any type of neural network can be trained with the help of the so-called backpropagation algorithm. Fundamental components of backpropagation are a neural network topology, a set of training data $(\mathbf{X}, \mathbf{Y})$ consisting of samples $\mathbf{X}$ and their labels $\mathbf{Y}$ as well as a so-called loss function $\mathcal{L}$. The latter defines the optimization objective, which commonly measures the deviation of the models output for every training sample from the desired target value $\in \mathbf{Y}$. Backpropagation devises a parameter-update strategy moving backwards through the network topology by making extensive use of the chain-rule for computing mathematical derivatives. Essentially, given a sample and its label, the derivative of the loss $\mathcal{L}$ can be computed with respect to every parameter of the ANN:

$$\frac{\partial}{\partial w}\mathcal{L}(f_L(x_i), y_i) = \frac{\partial}{\partial w}\mathcal{L}(f_L(f_{L-1}(...f_1(x_i))), y_i) \tag{2.6}$$

The extensive usage of the chain-rule allows to compute the derivative of the loss w.r.t. a parameter $w_{l,j}$ of layer l as a chain of previous derivatives:

$$\frac{\partial}{\partial w_{k,j}}\mathcal{L}(f_L(x_i), y_i) = \frac{\partial}{\partial f_L(x_i)}\mathcal{L}(f_L(x_i), y_i) \cdot \frac{\partial}{\partial f_{L-1}(x_i)}f_L(x_i) \cdot ... \cdot \frac{\partial}{\partial w_{k,j}}f_k(x_i) \tag{2.7}$$

To train a model, training data is propagated through the network and the respective loss is computed. The derivatives w.r.t. all weights are computed via backpropagation. Actual parameter updates depend on the chosen optimization algorithm, commonly a first-order gradient-descent technique. For large ANNs and large training datasets, processing and computing the derivatives for the weights from all training data in a single pass becomes intractable. Stochastic mini-batch gradient descent is widely adopted to iteratively update the models' weights from random subsets of the training data until convergence.

2.2.3 Convolutional Neural Networks

In the areas of Computer Vision and medical image analysis, input data usually lies in a high dimensional cartesian space. Under such conditions, vanilla ANNs pose infeasibly high demands on computational resources because of the large number of weights resulting from the dense neuron connectivity. For DL on image data, great advances have been made with the introduction of a specialized building block, the so-called *Convolutional Layer* [57]. This building block heavily reduces the number of parameters in ANNs and explicity captures geometric information. Most of the success stories in image classification, object detection, semantic segmentation, face or hand-writing recognition depend on neural networks composed of such layers.

Similar to the Perceptron, the Convolutional Layer has a biological analogue. The usage and stacking of convolutional layers, resulting in a Convolutional Neural Network (CNN), is inspired by the animal visual cortex. As such, the visual cortex can be understood as a collection of spatial filters that slide across their input and activate on occurrence of distinct visual patterns within the sliding window, the so-called local receptive field.

In the context of ANNs, such a spatial filter can be modelled as a set of neurons with sparse connectivity to a portion of the actual input. The connectivity pattern is replicated across the entire input with a certain striding and shared weights among all replicas. From a signal processing perspective this topology resembles the concept of discrete convolution of input data with a translation invariant filter, whose coefficients are defined by the aforementioned shared weights of the spatially replicated neurons. CNNs commonly comprise multiple, consecutive banks of such filters, often composed with non-linear activation functions and down-sampling operations. During optimization, the respective filter weights are tuned to respond maximally to task-specific patterns. More precisely, a layer l consisting of k filters $\mathcal{K} \in \mathbb{R}^{P \times Q \times C}$ operating on a C-channel input tensor $\mathcal{T}_{l-1} \in \mathbb{R}^{H \times W \times C}$ yields a set of k so-called feature-maps $\mathcal{T}_l \in \mathbb{R}^{H \times W \times k}$, each map being the result of a single convolutional kernel from that filter bank as the result of the convolution operation. Let K be a single filter or *kernel*, then the discrete convolution operation in the context of CNNs is defined as:

$$\mathcal{T}_l(x, y) = (\mathcal{T}_{l-1} * \mathcal{K})(x, y) = \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{c=1}^{C} \mathcal{T}_{l-1}(x - \frac{P}{2} + p, y - \frac{Q}{2} + q, c) \mathcal{K}(p, q, c) \tag{2.8}$$

In CNNs, convolution operations are commonly combined with subsampling operations such as max- or average-pooling. Alternating pooling and convolutional kernels kept at constant size allows CNNs to first to learn to respond to local cues, and then to increasingly more global patterns the deeper data enters the network (see Fig. 2.4).

2.2.4 Classification with Supervised Deep Learning

Supervised DL emerged as a powerful tool for classification tasks. Under the premise that pairs $(\mathbf{x}_i, y_i)$ of data $\mathbf{x}_i$ with class label y_i are available and abundant, ANNs are able to learn non-linear mappings from the data $\mathbf{x}$ to their label y. The learning is driven by gradually and iteratively optimizing the network's weights using SGD with backprop with the goal to minimize the training loss function over many iterations. For multi-class classification tasks, a widely used objective is the cross-entropy loss

$$\mathcal{L}(f_L(x_i), y_i) = -\sum_{c}^{C} \log(f_L(x_i)^c) \cdot y_i^c \tag{2.9}$$

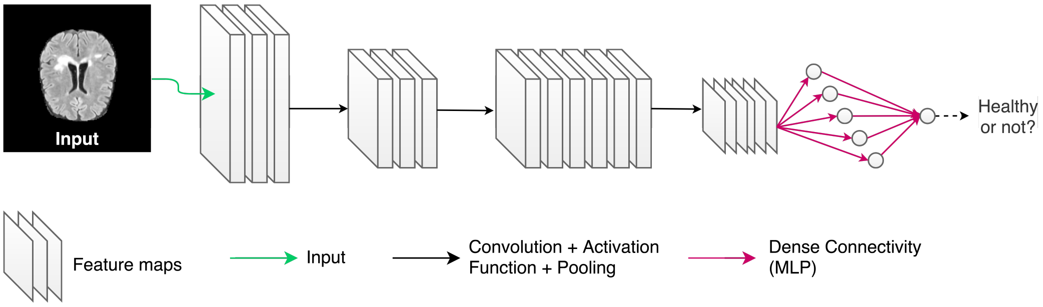


Fig. 2.4 Illustration of LeNet-5, one of the first CNNs proposed for recognizing human handwriting of single digits.

where C denotes the set of classes, $f_L(x_i)^c$ is the predicted probability that x_i is of class c and y_i^c is the ground-truth class label, i.e. a binary indicator which equals 1 if x_i is known to be of class c and 0 otherwise.

2.2.5 Segmentation with Supervised Deep Learning

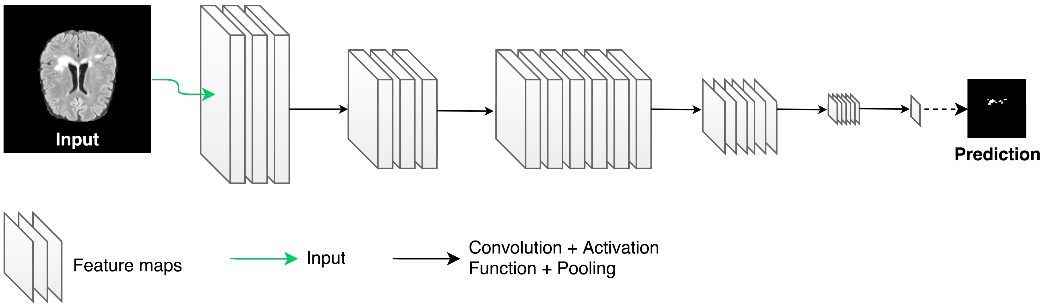


Fig. 2.5 Difference between CNNs and Fully Convolutional Networks (FCNs).

The segmentation of medical images (or image data in general) can be seen as a special classification task where every pixel (or voxel in case of volumetric data) is assigned a class label. Supervised DL has notably pushed the frontiers in various medical segmentation tasks. First successes primarily relied on patch-based classification; The approaches cropped potentially overlapping regions of interest, so-called image patches, from the data and classified all of them separately before aggregating all results into a final segmentation image. This admittedly laborious technique was soon sequeled by the much more efficient and effective Fully Convolutional Nets (FCN) [60]. FCNs are completely devoid of any densely connected neurons and instead only comprise convolutional kernels and pooling operations. Additionally, FCNs do not solely contract data to a lower-dimensional feature representation, but are enhanced by an expanding path-way. This enhancement allows to map from the low-dimensional latent features back to a space with dimensionality similar or equal to the model's input. The upsampling is effectively learned with the help of so-called de-convolutional or transpose-convolutional layers.

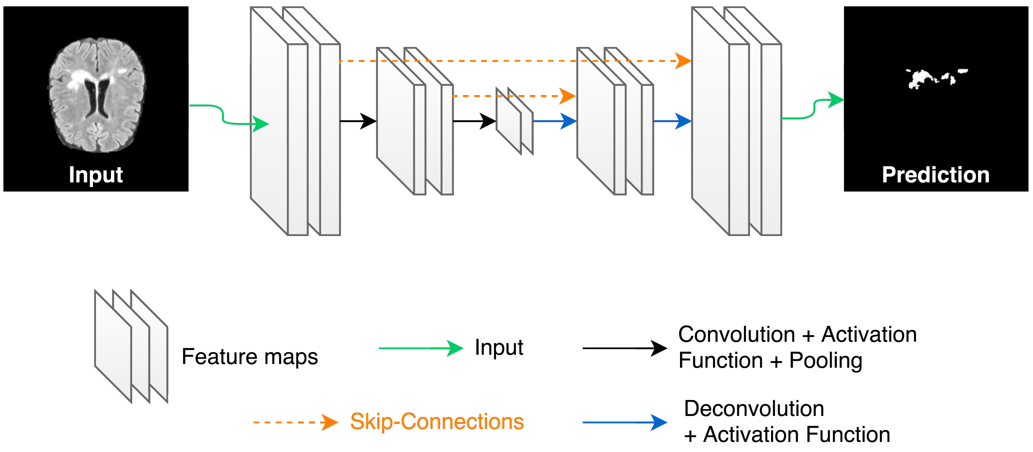

Input
Prediction
Feature maps
Input
Convolution + Activation Function + Pooling
Skip-Connections
Deconvolution + Activation Function

Fig. 2.6 Illustration of a U-Net-like segmentation network, one of the ground-breaking architectures for medical image segmentation. An encoder extracts increasingly global information from input data; The subsequent decoder network produces a segmentation with dimensionality similar or equal to the input data from the extracted information. Skip-connections between corresponding layers of the encoder and decoder allow to incorporate multi-scale, geometrical features into the segmentation process.

A particularly prominent representative of such FCNs in the medical domain is the so-called U-net [77], which further introduced skip-connections between corresponding layers of the contracting and the expanding path to provide its kernels with features extracted at various scales (see Fig. 2.6). This approach set the state-of-the-art in various medical segmentation tasks and has seen a wide array of extensions. With the V-Net [67] and 3D U-Net [31], the concept has been successfully transfered to volumetric data such as CT or MRI.

Analogue to CNNs, FCNs may be optimized with the help of the cross-entropy loss. Given an FCN architecture with input $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$ and a network output $f(\mathbf{x}) \in \mathbb{R}^{H \times W \times C}$ of the same dimensionality, an overall loss can be aggregated over all pixels (or voxels) of the multi-dimensional tensors $\mathbf{x}$ and $f(\mathbf{x})$. In medical imaging, where pathologies or anomalies are often less frequent than healthy surrounding anatomy, the optimization of FCNs introduces an additional challenge. Under class-imbalance, it is important to ensure that the model does not converge to a trivial solution where it classifies each pixels with the majority class. Special loss functions like the DICE-loss [67] aim to overcome this problem by implicitly equalizing the gradients of the classes. Alternative approaches rely on explicit weighting heuristics like median frequency balancing [36] which weights loss gradients of different classes.

2.3 Pathology-specific Brain Lesion Detection & Segmentation

In MR imaging of the brain, the majority of DL-research did not consider anomaly detection and segmentation as an outlier-detection problem. Instead, most of proposed methods were devoted to solving detection and segmentation tasks via explicit differentiation between normal anatomy and specific pathologies in different modalities and parts of the anatomy.

The distinct task of segmenting brain tumors from surrounding healthy anatomy has seen a great number of DL-driven contributions. A driving force is the Brain Tumor Segmentation Challenge “BraTS”, hosted annually since 2012. Each year, a vast array of new and improved methods for automatic brain tumor segmentation have entered the challenge and set new standards. Since the challenge’s launch, supervised DL-based approaches have been among the top performers. The first successes were reported with patch-based 2D[61, 71, 75, 105], which were then superseded by fully convolutional 2D [35, 43] and 3D networks [25, 50]. The majority of methods relied on multi-modal input: Zikic et al. [105] made an early attempt towards brain tumor tissue delineation with the help of shallow CNNs on multi-modal data. Lyksborg et al. [61] proposed a cascade of ensembles of cross-section-specific 2D patch-based CNNs and a subsequent ensemble which segments different tumor compartments. Rao et al. [75] combined modality-specific patch-based 2D CNNs with a Random Forest classifier; Pereira et al. [71] used deeper CNNs with multi-modal data and achieved top ranks in the BraTS 2013 and 2015 challenge. Dvorak et al. [35] employed 2D FCNs for structured label prediction in multi-modal image MR data. On mono-modal data, Havaei et al. [43] tailored a cascade of two separately trained FCNs with the goal of refining false positive predictions from the first network. All previous methods rely on 2D filters and thus do not exploit the natural 3D information provided in brain MRI. 3D convolutional kernels seem very appealing, but incur high computational demands. To overcome this, Kamnitsas et al. [50] proposed an efficient multi-scale 3D CNN with a fully connected Conditional Random Field (CRF) refinement for accurate brain lesion segmentation. [25] compared this approach to a variety of other multi-resolution 3D FCN architectures.

The challenging task of MS lesion segmentation has seen fewer focus, but the difficulty of the task is incomparably higher. An early DL-driven attempt has been made by Birenbaum et al. [20], who used CNNs on multi-view cross-sectional patches for longitudinal MS lesion segmentation. Valverde et al. [96] proposed a two-step cascaded 3D CNN, where the second model reduces the false positives produced by the first one. Prieto et al. [73] examined large deep neural networks for MS lesion segmentation; Roy et al. [80] employed a two-stage FCN with modality-specific filter-banks on 2D slices from multi-modal MR data to segment WML. Vaidya et al. [95] also employ 3D CNNs, but focus on MS lesion segmentation from longitudinal data.

2.4 Cross-Domain Brain Lesion Segmentation

MR images may vary greatly in terms of their intensity distribution and imaging characteristics across vendors and acquisition sites. In comparison to CT, MR intensities cannot be mapped to physical scales such as the Houndsfield Scale and are thus not clearly associated to specific tissue types. For supervised DL, this poses difficulties when it comes to analyzing images from scanners not present in the training data distribution. Often, performance drops or unpredictable behavior are witnessed when neural networks are applied to data which exhibit distribution shifts (i.e. data which comes from a slightly different domain). More labeled training data that covers the gamut of variabilities would be helpful, but is rarely readily available.

2.4.1 Semi-Supervised Methods for Domain Adaptation

Two different paradigms offer frameworks to overcome these limitations and enhance CNN's generalization capabilities across different domains. Semi-supervised DL provides tools for leveraging limited quantities of annotated training samples and arbitrary amounts of unlabeled training data. Domain adaptation is a DL-paradigm generally targeted at making models perform well on data from a different domain—the target domain—, whose distribution might differ from the training data—the source domain. Some semi-supervised DL frameworks may be used for domain adaptation as well.

Prior to the advent of FCNs, semi-supervised methods utilized embedding techniques and prior knowledge about the origin of data [100], more elaborate graph embeddings [101] or statistical pseudo-labelling approaches[59] to learn domain invariant latent feature representations in the neural network. A specifically tailored architecture with a ladder-like topology has also been crafted by Rasmus et al. [76].

2.4.2 Contribution: Semi-Supervised Learning for Fully Convolutional Networks

Aforementioned semi-supervised methods were primarily developed for traditional CNNs. However, fully-convolutional architectures such as the U-Net[77] have rapidly emerged as the de-facto standard for efficient medical image segmentation, and at that time FCNs lacked an adequate framework for semi-supervised learning, and domain adaptation in particular. In [10], the auxiliary embedding task proposed by Weston et al. [100] is adopted and lifted to fully convolutional architectures at the example of the U-Net in the challenging task of MS lesion segmentation. To keep the optimization tractable, a statistical feature sampling and embedding scheme was developed. This so-called "Random Feature Embedding" leverages prior knowledge in the form of a graph adjacency matrix and feature extraction from the high dimensional tensor of FCNs to learn domain invariant feature representations for features which share graph adjacency, and enforces more distinctive decision boundaries between non-adjacent feature vectors. An extensive ablative study on different feature sampling strategies and types of priors showed considerable improvements over the completely supervised baseline.

Contributions | The author of this thesis was responsible for the main idea of lifting auxiliary embedding tasks to multi-dimensional tensors with the help of Random Feature Embedding, the implementation of the proposed algorithms and frameworks, the experimental validation and the writing of the manuscript. Shadi Albarqouni contributed with discussions and feedback on the main ideas of the paper, experimental design and evaluation, and was involved in proof-reading. Nassir Navab contributed with discussions and feedback on the main ideas of the paper and was involved in proof-reading.

Copyright statement | This contribution was originally published as: "**Christoph Baur**, Shadi Albarqouni, and Nassir Navab. "Semi-supervised deep learning for fully convolutional networks." *In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 311-319. Springer, Cham, 2017*.".

Reprint Denied

The reprint of this publication was rejected on open-access platforms. The publication can be found at (`https://link.springer.com/chapter/10.1007%2F978-3-319-66179-7_36`). The details are provided below.

Publication [10]

Authors: Christoph Baur, Shadi Albarqouni, Nassir Navab
Title: “Semi-supervised deep learning for fully convolutional networks”
In: International Conference on Medical Image Computing and Computer-Assisted Intervention

DOI: 10.1007/978-3-319-66179-7_36

2.4.3 Towards Unsupervised Domain Adaptation

In subsequent work on domain adaptation, Kamnitsas et al. [49] developed a completely unsupervised scheme, exemplified on brain lesion segmentation. The method relies on the adversarial training concept known from GANs (3.2.3) to learn domain-invariant feature representations. As a completely unsupervised domain adaptation method, no segmentation ground-truth is required for the target domain data. A domain discriminator network is jointly optimized with the actual segmentation network and exploited as a proxy loss to minimize the discrepancy between feature representations of both the source and target domain.

Unsupervised Deep Learning for Anomaly Detection in Brain MRI

3

3.1 Definition of Unsupervised Anomaly Detection

Supervised DL explicitly learns to differentiate between what may be considered normal or *background* and instances of what constitutes an anomaly. In contrast, UAD affiliates with terms such as “outlier detection” or “out-of-distribution detection”, i.e. UAD does not make assumptions about the notion of anomalies. Unsupervised methods either i) make no assumptions about the data at all and must inherently learn about the likelihood of samples from sets which may or may not contain anomalous instances, or ii) do not have any knowledge about what precisely depicts an anomaly, but explicitly model a normative distribution of healthy anatomy. The first family of approaches may be primarily linked to dimensionality reduction and clustering techniques. In this thesis, the majority of presented methods and contributions focus on the second case, which is facilitated by recent advances in the field of generative modeling.

Similar to anomaly detection in a supervised context, UAD may be formulated as a per-sample instance classification task, a localization task or even as an Unsupervised Anomaly Segmentation (UAS) problem.

3.2 Unsupervised Deep Representation Learning & Deep Generative Models

In brain imaging, UAD has seen a wide array of methods, ranging from clustering-techniques based on unsupervised representation learning with AEs to the recent wave of distribution modeling with deep generative modeling frameworks such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs) and hybrids thereof. In the following, a conceptual overview and distinction is given on these different frameworks.

3.2.1 Autoencoders

AEs are a specific class of ANNs and a crucial component of the “unsupervised” learning realm. They consist of a so-called encoder-network $\mathrm{Enc}_\theta(\mathbf{x})$ with parameters θ, which maps high-dimensional input $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$ to a lower-dimensional latent space $\mathbf{z} \in \mathbb{R}^d$, and a subsequent decoder network $\mathrm{Dec}_\phi(\mathbf{z})$ with parameters ϕ, which is tasked to recover or *reconstruct* the input

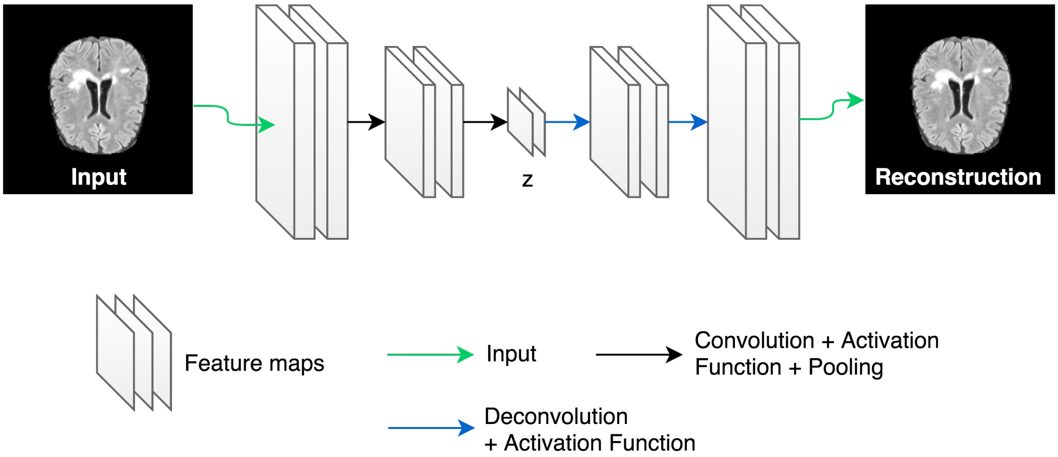

Input
z
Reconstruction
Feature maps
Input
Convolution + Activation Function + Pooling
Deconvolution + Activation Function

Fig. 3.1 A convolutional AE with a spatial bottleneck.

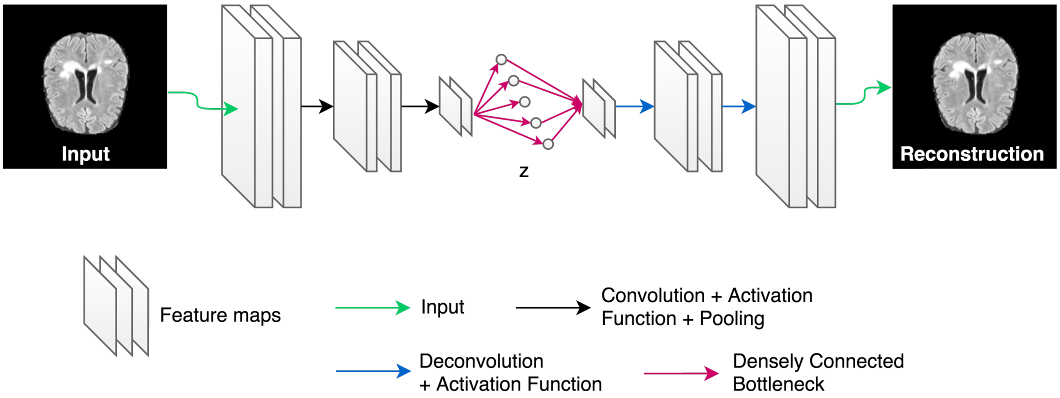

Input
z
Reconstruction
Feature maps
Input
Convolution + Activation Function + Pooling
Deconvolution + Activation Function
Densely Connected Bottleneck

Fig. 3.2 A convolutional AE with a dense bottleneck.

$\mathbf{x} \in \mathbb{R}^{H \times W \times C}$ from the lower dimensional representation $\mathbf{z} \in \mathbb{R}^d$ [81] (see Fig. 3.2). AEs may be used for dimensionality reduction, clustering or lossy data compression. Moreover, AEs have been employed for pre-training ANNs [18] using unlabeled data to speed up convergence during successive training of a supervised model; Further, they have been used for signal denoising tasks [98].

AEs are unsupervised as their training is free of labeled data. Their optimization involves the minimization of a reconstruction loss $\mathcal{L}_{Rec}$ between the training input $\mathbf{x}$ and its reconstruction $\hat{\mathbf{x}}$:

$$\underset{\phi,\theta}{\operatorname{argmin}} \mathcal{L}_{Rec}^{\phi,\theta}(\mathbf{x}, \hat{\mathbf{x}}) = \ell_k(\mathbf{x}, \hat{\mathbf{x}}) \tag{3.1}$$

, where ℓ_k denotes a differentiable norm and k defines the type of norm to be computed. Typical choices are $k = 1$ or $k = 2$ for the ℓ_1 or ℓ_2-norm, respectively.

AEs come in many forms and variations. Early work used DNNs [81] as illustrated above. On imaging data, convolutional AEs are more common [64]. The bottleneck of convolutional AEs may also be tensor-shaped, yielding so-called spatial AEs that bring along improved reconstruction capabilities due to the preservation of geometry in the compressed representation (see Fig. 3.1).

3.2.2 Variational Autoencoders

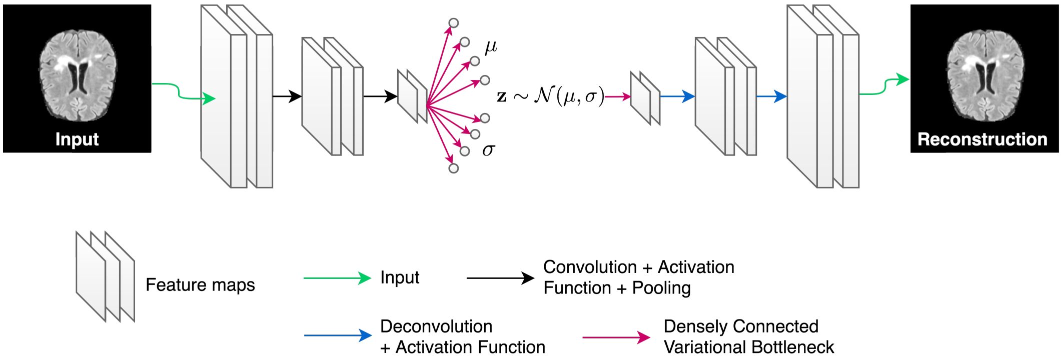

Input
μ
$\mathbf{z} \sim \mathcal{N}(\mu, \sigma)$
σ
Reconstruction
Feature maps
Input
Convolution + Activation Function + Pooling
Deconvolution + Activation Function
Densely Connected Variational Bottleneck

Fig. 3.3 An illustration of the VAE. An encoder network regresses the parameters of an approximate posterior distribution, i.e. the mean μ and variance σ of $q(\mathbf{z}|\mathbf{x})$. A sample $\mathbf{z}$ drawn from this distribution is fed into a decoder network, which tries to reconstruct or generate a sample from the modeled data distribution.

The VAE [54] is a specific type of generative model that embeds bayesian variational methods into an AE framework. As a generative model, the VAE tries to capture the underlying distribution $p(\mathbf{x})$ of training data. In the VAE this is facilitated with the help of a surrogate distribution with lower dimensionality, i.e. an approximate posterior distribution $q(\mathbf{z}|\mathbf{x})$. This approximate posterior is parameterized by the encoder network and trained to follow a prior distribution of choice. By sampling from the prior distribution and passing the sample through the decoder network, data from the modeled distribution can be generated (see Fig. 3.3 for a depiction of the VAE).

To approximate the underlying distribution $p(\mathbf{x})$ of data $\mathbf{x} \sim \mathbf{X}$, deep generative models typically leverage a training data-set $\mathbf{X} = \{\mathbf{x}^{(i)}\}_{i=1}^{N}$ of cardinality N to learn features and their latent relationships directly from data. In [54], the authors hypothesize that finding an explicit model for $p(\mathbf{x})$ is intricate and instead propose a latent variable model with a low-dimensional random variable $\mathbf{z} \in \mathbb{R}^d$ that is supposed to capture and disentangle the essential latent factors in the data. The posterior distribution of $\mathbf{z}$ can then be determined using the so-called Variational Bayes framework. This framework allows to approximate the true posterior distribution $p(\mathbf{z}|\mathbf{x})$, which usually cannot be solved in a tractable way due to the required computation of the marginal likelihood $p(\mathbf{x}) = \int p(\mathbf{x}, \mathbf{z}) d\mathbf{z}$ via integration over all $\mathbf{z}$. A surrogate distribution $q(\mathbf{z}|\mathbf{x})$ from a known family of distributions is used instead and a lower bound to the log-likelihood of $p(\mathbf{x})$ is maximized:

$$\begin{aligned}\log p(\mathbf{x}) &= \log \int_z p(\mathbf{x},\mathbf{z})dz \\ &= \log \int_z p(\mathbf{x},\mathbf{z})\frac{q(\mathbf{z}|\mathbf{x})}{q(\mathbf{z}|\mathbf{x})} \\ &= \log \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}\left[\frac{p(\mathbf{x},\mathbf{z})}{q(\mathbf{z}|\mathbf{x})}\right] \\ &\overset{\text{Jensen's inequality}}{\geq} \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}\left[\log\frac{p(\mathbf{x},\mathbf{z})}{q(\mathbf{z}|\mathbf{x})}\right] \\ &= \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}\left[\log\frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{q(\mathbf{z}|\mathbf{x})}\right] \\ &= \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}\left[\log p(\mathbf{x}|\mathbf{z})\right] - \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}\left[\log\frac{p(\mathbf{z})}{q(\mathbf{z}|\mathbf{x})}\right] \\ &= \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}\left[\log p(\mathbf{x}|\mathbf{z})\right] - \mathcal{D}_{KL}(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) = ELBO\end{aligned} \tag{3.2}$$

From Eq. 3.2, it can be seen that maximizing the log-likelihood of the data distribution $p(\mathbf{x})$ can be approximated by maximizing the so-called Evidence Lower-Bound (ELBO), a lower bound to the marginal likelihood that is free of any intractable terms. This lower bound is derived using Jensen's inequality [47]. As a result, the objective involves the simultaneous optimization of the variational parameters of $q(\mathbf{z}|\mathbf{x})$ and the generative parameters of $p(\mathbf{x}|\mathbf{z})$ via minimization of the KL-Divergence between $q(\mathbf{z}|\mathbf{x})$ and the prior $p(\mathbf{z})$ as well as the maximization of the expectation term:

$$\begin{aligned}\mathcal{L}_{VAE} &= -\mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[log(p(\mathbf{x}|\mathbf{z}))] + \mathcal{L}_{prior}\mathcal{D}_{KL}(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) \\ &= \mathcal{L}_{rec} + \mathcal{L}_{prior}\mathcal{D}_{KL}(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))\end{aligned} \tag{3.3}$$

In the VAE, an encoder network $\mathrm{Enc}_\theta(\mathbf{x})$ with network parameters θ is used to learn to regress the parameters of $q(\mathbf{z}|\mathbf{x})$, i.e. the mean μ and variance σ of the approximate posterior. Similarly, the decoder network $\mathrm{Dec}_\phi(\mathbf{z})$ with network parameters ϕ determines the parameters of $p(\mathbf{x}|\mathbf{z})$, with $\mathbf{z} \sim q(\mathbf{z}|\mathbf{x})$. Since the optimization of network parameters requires differentiable components, the stochastic nature of the ELBO must be made differentiable to facilitate gradient flow. Kingma et al. [54] propose the so-called reparameterization trick which replaces the stochasticity of the ELBO-terms by a deterministic function parameterized with another random variable. The overall loss function comprises a reconstruction loss term between the input $\mathbf{x}$ and the reconstruction $\hat{\mathbf{x}} \sim \mathrm{Dec}(\mathbf{z}), \mathbf{z} \sim \mathrm{Enc}(\mathbf{x})$, and a regularizing KL-Divergence term which ensures that $q(\mathbf{z}|\mathbf{x})$ approximates the chosen prior $p(\mathbf{z})$ (typically a multivariate normal distribution). From a practical perspective, this regularizer ensures locality and smoothness in the latent space for similar $\mathbf{x}$, while the reconstruction loss term ensures proper reconstruction of the input samples.

The VAE has seen a wide array of extensions. The Adversarial Autoencoder (AAE) [62], inspired by the GAN training, leverages an adversarial network as a proxy to the KL-divergence regularization term for improved disentanglement of the latent space and the ability to match it to arbitrary priors $p(\mathbf{z})$. The Gaussian Mixture Variational Autoencoder (GMVAE) [33] even replaces the uni-modal posterior $q(\mathbf{z}|\mathbf{x})$ by a mixture of gaussian for higher expressive capabilities and improved latent structure.

3.2.3 Generative Adversarial Networks

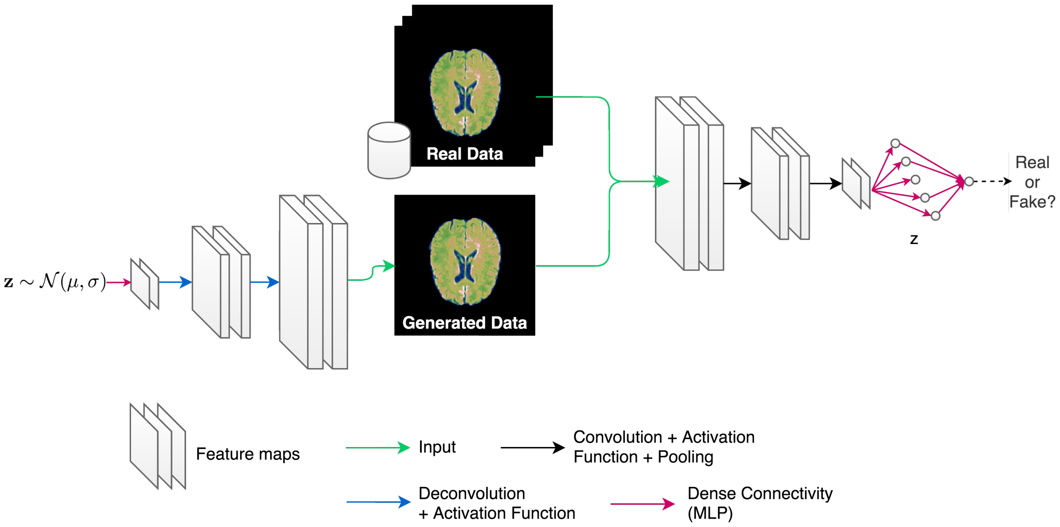


Fig. 3.4 An illustration of the GAN with brain MR data as an example.

GANs [40] describe a different, unique approach towards deep generative modeling. Instead of embedding the modeling of a distribution inside a feed-forward reconstruction task (see Sub-section 3.2.2 on VAEs), GANs leverage a two-player mini-max game between two ANN opponents to learn a parametric, generative model which maps from a prior distribution $p(\mathbf{z})$ to data samples. More precisely, a generator network $\mathrm{G}(\mathbf{z})$ is optimized to map from random samples $\mathbf{z} \sim p(\mathbf{z})$ to data $\hat{\mathbf{x}}$. Another discriminator network $\mathrm{D}(\mathbf{x})$ is tasked to differentiate between real samples $\mathbf{x_i}$ from the training data-set $\mathbf{X}$ and generated data $\hat{\mathbf{x}}$ coming from G. The networks constitute adversaries as they are trained in an alternating manner while G being solely trained through the "feedback" from D. The generator's parameters are iteratively adjusted such that D is likely to misclassify the generated input as real. As training progresses and the discriminator D improves, G's generation capabilities evolve as well. GANs have shown the capability to synthesize data at outstanding levels of realism and therefore have heavily influenced other areas of DL-research [53] such as image-to-image translation, super-resolution, denoising, image reconstruction, segmentation and even image classification. The general framework is depicted in Fig. 3.4. The training objective may be formulated as:

$$\min_{\mathrm{G}} \max_{\mathrm{D}} V(\mathrm{D}, \mathrm{G}) = \mathbb{E}_{\mathbf{x} \sim p_{data}(\mathbf{x})}[log(\mathrm{D}(\mathbf{x}))] + \mathbb{E}_{\mathbf{z} \sim p_z(\mathbf{z})}[1 - log(\mathrm{D}(\mathrm{G}(\mathbf{z})))] \tag{3.4}$$

Notably, the output of D is simply a sigmoid-activated "probability" of its belief of the input sample being real. The goal of the training is to find a model D which a) maximizes the expectation for samples drawn from the real data distribution to be correctly classified. At the same time, the objective strives for a G which makes D think that $\mathrm{G}(\mathbf{z})$ is real such that $D(G(\mathbf{z}))$ yields 1 and the whole term becomes minimal. The optimal solution of this value function $V(\mathrm{D},\mathrm{G})$ should lead to a so-called Nash Equilibrium, i.e. the optimal outcome of the "game" where no player has an incentive to deviate from his chosen strategy after considering the opponent's choice. The fact that the framework is simply a combination of two neural networks allows for training via backpropagation. Goodfellow et al. turn the optimization for $V(\mathrm{D},\mathrm{G})$ into a two-step strategy, where in each iteration, they update D and G separately:

$$\max_{\mathrm{D}} \mathbb{E}_{\mathbf{x}\sim p_{data}(\mathbf{x})}[log(\mathrm{D}(\mathbf{x}))] + \mathbb{E}_{\mathbf{z}\sim p_z(\mathbf{z})}[1 - log(\mathrm{D}(\mathrm{G}(\mathbf{z})))] \tag{3.5}$$

$$\min_{\mathrm{G}} \mathbb{E}_{\mathbf{z}\sim p_z(\mathbf{z})}[1 - log(\mathrm{D}(\mathrm{G}(\mathbf{z})))] \tag{3.6}$$

GANs entail a variety of peculiarities: i) G is trained via backpropagation from D without requiring any labels; ii) the authors proved that the GAN training scheme minimizes the Jensen-Shannon divergence between the distributions of real and generated data, provided the models D and G have sufficient capacity. In practice, training GANs this way is daunting and brings along some challenges of its own. During training, GANs may easily collapse towards generating only data from a single mode of the data distribution for the majority of values of $\mathbf{z}$ ("mode collapse"), and the sigmoid-cross entropy loss of D introduces a vanishing gradient problem. Early during training, it is very difficult to train G, as D can easily differentiate between real and generated samples, which leads to low magnitude gradients and thus no significant updates to G. Extensions to GANs such as the Deep Convolutional Generative Adversarial Network (DCGAN) [74], turn the originally DNN architecture into a deconvolutional neural network, allowing the synthesis of image-data with higher resolution. The carefully designed architecture further eases and stabilizies the training. Other work identified the loss function as the culprit for mode collapse, and suggests to use f-divergences [69], the Wasserstein-divergence [4] or the least-squares loss [63] for improved training stability and image quality. A variety of works introduce additional auxiliary networks and loss functions to enforce constraints on the latent feature space of the discriminator [28] or the latent code $\mathbf{z}$ [26]. Others even suggest to formulate the optimization problem as an energy minimization problem [103] rather than using a probabilistic formulation and deduce that it allows for more stable training and synthesis of much higher resolution images. Recently, progressive GAN growing has shown to produce very sharp and realistic high-resolution images [52] with successful transfer to the medical domain [8]. Numerous other essential frameworks have been derived from the original GAN formulation for tackling loosely related tasks.

Conditional GAN | The Conditional Generative Adversarial Network (cGAN) [68] enhances the generator's input by a supervisory signal to give explicit control on the data synthesis process. In addition to the random noise $\mathbf{z}$, G is jointly provided with some prior information $\mathbf{c}$. This prior knowledge is also passed to D jointly with the respective real or generated data during training. The optimization is slightly altered to:

$$\min_{\mathrm{G}} \max_{\mathrm{D}} V(\mathrm{D},\mathrm{G}) = \mathbb{E}_{x\sim p_{data}(\mathbf{x})}[log(\mathrm{D}(\mathbf{x}|\mathbf{c}))] + \mathbb{E}_{\mathbf{z}\sim p_z(\mathbf{z})}[1 - log(\mathrm{D}(\mathrm{G}(\mathbf{z}|\mathbf{c})))] \tag{3.7}$$

The provisioning of conditional information stabilizes the GAN training and enhances the synthesis quality.

Pix2Pix | The so-called Pix2Pix framework [45] constitutes a member of the cGAN-family targeted to image-to-image translation & style-transfer tasks. It combines a U-net-like generator network for the translation with a fully-convolutional discriminator network to facilitate the discrimination between real and generated high-resolution data. The conditional information is given by the input images to the U-net, and random noise is injected into the generator's bottleneck. For training, the framework requires paired data in both source and target style. The U-Net is optimized to transform samples with source style to the target style with the help of the discriminator, which in turn learns to distinguish between packages of real source and target style data, or real source and generated target style data, respectively. To stabilize the training, an ℓ_1 loss between the generators' output and the real target is employed. The skip-connections in the U-net-like generator play a vital role in the preservation of global coherence of style-transfered samples.

CycleGAN | While Pix2Pix requires paired training data, the CycleGAN [104] learns a style-transfer between two distributions without explicit sample pairs. The framework chains two different cGANs and introduces a cycle-consistency loss to stabilize their training. The first generator G learns to map from distribution $\mathbf{X}$ to distribution $\mathbf{Y}$ with the help of a discriminator for domain $\mathbf{Y}$ data, and the second generator F is optimized to establish the reverse mapping from distribution $\mathbf{Y}$ to distribution $\mathbf{X}$ with the help of a discriminator for domain $\mathbf{X}$:

$$L(\mathrm{G},\mathrm{F},\mathrm{D}_{\mathbf{X}},\mathrm{D}_{\mathbf{Y}}) = L_{GAN}(\mathrm{G},\mathrm{D}_{\mathbf{Y}},\mathbf{X},\mathbf{Y}) + L_{GAN}(\mathrm{F},\mathrm{D}_{\mathbf{X}},\mathbf{Y},\mathbf{X}) + \lambda L_{cycle}(\mathrm{G},\mathrm{F}) \tag{3.8}$$

where

$$L_{cycle}(\mathrm{G},\mathrm{F}) = \mathbb{E}_{x\sim p_{data}(\mathbf{X})}[||\mathrm{F}(\mathrm{G}(\mathbf{x})) - \mathbf{x}||_1] + \mathbb{E}_{\mathbf{y}\sim p_{data}(\mathbf{Y})}[||\mathrm{G}(\mathrm{F}(\mathbf{y})) - \mathbf{y}||_1] \tag{3.9}$$

minimizes the ℓ_1-reconstruction error between an input from domain $\mathbf{X}$ (or $\mathbf{Y}$, respectively) and a complete cycle through the chained generators $\mathrm{F}(\mathrm{G}(\mathbf{X}))$ (or $\mathrm{G}(\mathrm{F}(\mathbf{Y}))$, respectively). As for GANs, optimization is performed in an alternating manner by switching between discriminators and generators, but also between the cycles.

3.2.4 Hybrids

Hybrids between GANs and VAEs also exist. Larsen et al. [56] proposed the *VAE-GAN* to combine the unique strengths of both methods. An encoder network Enc is trained to project input data onto the manifold $\mathbf{z}$ which is supposed to follow a prior distribution, from which a decoder Dec recovers the input. A discriminator D further learns to distinguish between actual input and reconstructed (or generated) data, and provides gradients to the decoder Dec

to produce compelling reconstructions. The VAE-components of this framework, composed of the encoder and decoder, stabilize the training and allow an efficient determination of where input data lies on the latent manifold, while the GAN-components (decoder and discriminator) facilitate the generation of realistic, crisp images. The training setup is given by:

$$\mathcal{L}_{VAEGAN} = \mathcal{L}_{rec} + \mathcal{L}_{prior} + \mathcal{L}_{GAN} \tag{3.10}$$

As a reconstruction loss term, the authors employ a Gaussian observation model on the feature space of the discriminator, i.e. they minimize the ℓ_2-distance of the latent feature representations of the input $\mathbf{x}$ and the reconstruction $\hat{\mathbf{x}}$:

$$\mathcal{L}_{rec} = -\mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[log(p(\mathbf{x}|\mathbf{z}))]p(\mathrm{D}(\mathbf{x})|\mathbf{z}) = \mathcal{N}(\mathrm{D}(\mathbf{x})|\,\mathrm{D}(\hat{\mathbf{x}}), \mathbf{I}) \tag{3.11}$$

The authors particularly emphasize that the discriminator D may be interpreted as a learned similarity metric which considers image features beyond the pixel-level. The latter constitutes a particular short-coming of the VAE, which optimizes for every single data coefficient separately and independently, potentially leading to blurry reconstructions. On the contrary, the discriminator explicitly encourages the generation of realistic features in synthetic or reconstructed data. With ALI [78] and BiGAN [34], two very similar settings have been explored. Both frameworks enhance the GAN by a third feature encoder network, which is tasked to map generated data back onto the manifold from where it has been generated. Opposed to the VAE-GAN, no explicit reconstruction loss component is needed. Instead, by discriminating jointly on both the data and the latent code, it is formally proven that the encoder reverts the generator and vice versa [34].

3.3 GAN-based Anomaly Detection

3.3.1 Concept and Related Work

Pioneering work which exploits unsupervised deep generative modeling for UAD was presented by Schlegl et al. [87]. The authors proposed to learn a normative distribution of healthy anatomy of the retina using healthy retinal OCT patches with the help of a DCGAN. As a result, the trained generator is only able to synthesize anomaly-free OCT images, as it has only seen anatomically normal OCT data during training. This circumstance is exploited in an anomaly detection framework developed on top of the DCGAN. To determine whether a query image $\mathbf{q}$ carries any anomalies, Schlegl. et al [87] first randomly sample a vector from $p(\mathbf{z})$, pass it through the generator network to generate a synthetic retinal patch and then iteratively move along the manifold $\mathbf{z}$ of their trained GAN to minimize an anomaly score between the generated data $\mathrm{G}(\mathbf{z})$ and the query image $\mathbf{q}$ with the help of the backpropagation algorithm. Notably, the parameters of the trained DCGAN, i.e. the generator and discriminator, remain unchanged. A rough delineation of the anomalies can be obtained by computing the residuals between the query data and the result of the iterative restoration of the query image.

For higher-resolution data, an aggregation of patches and their costly iterative inference is required, though. The aforementioned anomaly-score consists of a residual term r and a discrimination term d. The residual term r measures the Mean Squared Error between $\mathrm{G}(\mathbf{z})$ and $\mathbf{q}$, which should be minimized. The discrimination term operates on the latent feature representations produced by the discriminator network when feeding in $\mathbf{q}$ and $\mathrm{G}(\mathbf{z})$ and tries to minimize the ℓ_2-norm between the two feature representations, i.e. enforcing the synthesis of a sample which the discriminator can hardly identify as fake. The scalar sum of this anomaly-score also serves as the means to indicate the presence of any anomalies by thresholding it.

In recent follow-up work, i.e. the so-called f-AnoGAN [86], Schlegl et al. replace the iterative optimization by a single feed-forward mapping from the query-sample onto the manifold and exchange the DCGAN for the more robust Wasserstein-GAN. In a first training step, they model the distribution of healthy retinal anatomy using the Wasserstein-GAN. In a second step, while keeping the weights of the GAN fixed, they train an encoder network to map healthy data onto the manifold of the latent distribution $p(\mathbf{z})$.

3.3.2 Contribution: SteGANomaly: Inhibiting CycleGAN Steganography for Unsupervised Anomaly Detection in Brain MRI

To identify anomalies in brain MRI from residuals between query data and GAN-based reconstructions with great precision, it is vital that morphology is not altered by the GAN and reconstructions preserve a great level of important anatomical details. By default, GANs mimic a data distribution and as such do not provide these guarantees [15]. In SteGANomaly, instead of explicitly modeling healthy anatomy using standard GANs, anomaly detection is embedded into a GAN-based style-transfer tasks using a state-of-the-art unsupervised image-to-image translation technique (the CycleGAN) and recent discoveries of its steganographic properties. CycleGANs are generally known for their effective mappings between two distributions and their ability to recover almost perfect reconstruction of their input even from representations with lower entropy. A thorough examination of this phenomenon has revealed that information is not lost during the translation process as the CycleGAN learns to hide the essential information of the higher-entropy data in imperceptible high frequency components in the allegedly lower entropy distribution style, from which data can be reconstructed with great fidelity. The concept investigated in this contribution relies on mapping healthy anatomy from real brain MR scans to a lower-entropy, simulated distribution of healthy brain MRI and back to the input distribution. The original rationale behind this is that anomalous data will then be mapped to the simulated space in which anomalies are replaced by normal anatomy. Consequently, the anomalies then should have vanished in the reconstruction of the real data as well when completing cycle. This work shows that this is indeed possible, but only when preventing the CycleGAN from placing high-frequency steganographic information in the lower-entropy, intermediate representation. The resulting framework can handle high-resolution brain MR slices and allows to detect and delineate brain lesions from residuals between input and the completed cycle. In a variety of ablative studies, the superiority of this approach over other state-of-the-art methods is shown.

Contributions | The author of this thesis was responsible for the main idea of embedding anomaly segmentation into a CycleGAN-based style-transfer framework by establishing a reversible mapping between real and simulated data, the design of the proposed framework and the evaluation pipeline, planning of the experiments, data preparation and the writing of the manuscript. Together with Robert Graf, he jointly contributed with the idea of investigating and constraining the steganographic properties of the CycleGAN for suppressing anomalies. Robert Graf was further responsible for the implementation, experimental validation and also contributed to the writing of the manuscript. Benedikt Wiestler contributed with discussions and feedback on the main ideas of the paper, as well as with datasets and their preparation. Shadi Albarqouni contributed with discussions and feedback on the main ideas of the paper, experimental design and evaluation, and was involved in proof-reading. Nassir Navab contributed with discussions and feedback on the main ideas of the paper and was involved in proof-reading.

Reprint Denied

The reprint of this publication was rejected on open-access platforms. The publication can be found at (`https://link.springer.com/chapter/10.1007%2F978-3-030-59713-9_69`). The details are provided below.

Publication [12]

Authors: Christoph Baur, Robert Graf, Benedikt Wiestler, Shadi Albarqouni, Nassir Navab
Title: "SteGANomaly: Inhibiting CycleGAN Steganography for Unsupervised Anomaly Detection in Brain MRI"
In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II 23
DOI: https://doi.org/10.1007/978-3-030-59713-9_69

3.4 Autoencoder-based Anomaly Detection

3.4.1 Concept and related work

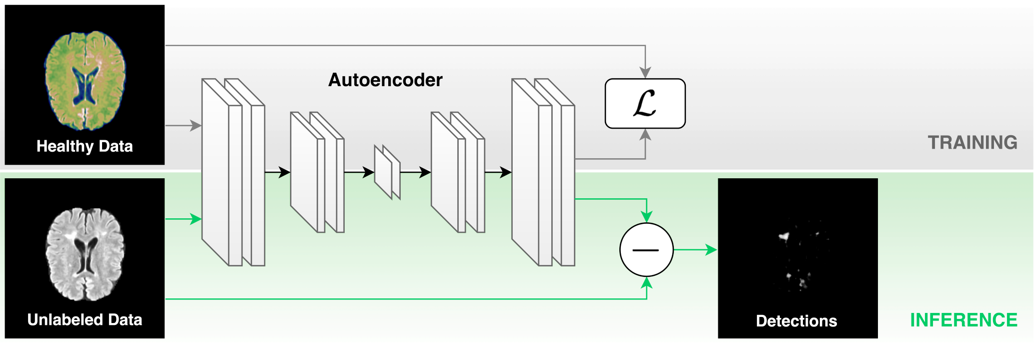


Fig. 3.5 Training and inference of AE-based models for UAS in image-space from reconstruction residuals (Source: [11]).

Upon successful training, GANs can generate crisp, realistic-looking image data from a distribution encoded in a user-defined manifold. However, the mapping between the manifold $\mathbf{z}$ and generated samples is not obvious. In the context of UAD, this makes the determination of a data-sample in the manifold $\mathbf{z}$ or an image-reconstruction of its healthy counterpart time-consuming. AEs are a simple and efficient alternative to GANs that have seen a vast number of applications to anomaly detection tasks ranging from one-class and out-of-distribution detection from its lower-dimensional manifold to pixel-wise anomaly detection in image-space [55]. In fact, in [84] it is shown that AEs learn well structured representations of normality and exhibit different activation behavior on abnormal samples. Outside the medical domain, a substantial amount of conceptual work was published with applications to toy datasets and anomaly detection and localization in videos [2, 42, 83]. In the medical field, Seeböck et al. [90] have been one of the first to employ AEs for UAD. They trained a convolutional AE from only healthy samples and combined it with a one-class SVM on the model's bottleneck to identify outlier samples from retinal OCT data.

On medical imaging data—and brain MRI in particular—a great deal of AE-based methods for UAS, i.e. voxel-based outlier detection in image-space, has recently appeared. Instead of classifying input samples as either normal or anomalous, these methods provide anomaly indications for every single image intensity value. The proposed methods can essentially be divided into three categories: i) reconstruction-based methods, ii) restoration-based methods and iii) gradient-based methods [11].

Reconstruction-based UAD | AEs [6, 15] and the generative VAEs [15, 30, 70, 106, 107] have been used to compress or explicitly model a normative distribution of healthy brain anatomy, respectively. Similar to the concept of Schlegl et al. [87], erroneous or inpainted reconstructions of anomalous data enable the discovery of anomalies from residuals between input and output of the models (see Fig. 3.5). The nature of AE-based models allows to directly project samples onto the lower-dimensional manifold $\mathbf{z}$ and to generate normal-looking reconstructions of abnormal data very efficiently in a single forward pass through the trained model. Thresholding of the resulting residuals yields pixel- or voxel-wise indications of anomalies directly in image-space. In [11, 15], different types and architectural designs of autoencoding models are investigated for UAS in brain MRI, and a combination of VAEs and GANs is proposed to improve both modeling capabilites and reconstruction fidelity. Chen et al. [30] investigate the effect of constraining the AAE, and [107] formulate the modeling and reconstruction of healthy anatomy as an inpainting task. Pawlowski et al. [70] employ Monte-Carlo (MC)-methods on top of VAE-based generative modeling of normal anatomy to obtain a reconstruction consensus for improved unsupervised brain tumor detection.

Restoration-based UAD|As exemplified in [87], restoration-based approaches try to explicitly alter samples in an iterative optimization to make them more similar to data from the actual modeled distribution. The ELBO in VAEs can be used as a backpropagation objective to transform input data while increasing its log-likelihood of being part of the prior distribution. You et al. [11] introduce this concept on VAEs and also lift it to the more expressive GMVAEs. Once converged, a residual image between input and restoration can be used to segment anomalies.

Gradient-based UAD|Unlike restoration-based methods, gradient-based approaches [106, 107] undertake only a single optimization step towards an objective like the ELBO. The computed gradient image acts a saliency map to detect regions with low likelihood, i.e. regions which constitute anomalies.

3.4.2 Contribution: Deep Autoencoding Models for Unsupervised Anomaly Segmentation in Brain MRI

This investigation of auto-encoding models is one of the first works in the field of DL-driven UAS in brain MRI and compares different types of AEs and generative models. AEs, VAEs and GANs are employed to reliably model the distribution of healthy brain anatomy of complete, high-resolution brain MR slices. In contrast to the previous patch-based approach from Schlegl et al. [87], a single feed-forward pass through the AE-models is sufficient to yield reconstructions of input data in which brain lesions are replaced by normal-looking tissue. As exemplified on real brain MR scans of subjects with MS, a simple pixel-wise comparison of the input and the respective reconstruction allows to spot and delineate anomalies such as White-Matter Hypterintensities. Not only different frameworks, but different choices for the degree of compression in the AE-models are investigated. Best segmentation performance is achieved with a geometry-preserving spatial bottleneck. Another contribution of this work is the combination of VAEs and GANs into the *Ano-VAEGAN*, an adaptation of the VAE-GAN [56] to combine the strengths of both generative modeling techniques into a fast, reliable framework for UAS.

Contributions|The author of this thesis was responsible for the main idea of leveraging AEs and VAEs for UAD, for the idea of combining VAEs and GANs into the Ano-VAEGAN, for the design and implementation of the proposed frameworks and the evaluation pipeline, planning of the experiments, data preparation and the writing of the manuscript. Benedikt Wiestler contributed with discussions and feedback on the main ideas of the paper, as well as with datasets and their preparation. Shadi Albarqouni contributed with discussions and feedback on the main ideas of the paper, experimental design and evaluation, and was involved in proof-reading. Nassir Navab contributed with discussions and feedback on the main ideas of the paper and was involved in proof-reading.

Reprint Denied

The reprint of this publication was rejected on open-access platforms. The publication can be found at (`https://link.springer.com/chapter/10.1007/978-3-030-11723-8_16`). The details are provided below.

Publication [15]

Authors: Christoph Baur, Benedikt Wiestler, Shadi Albarqouni, Nassir Navab
Title: “Deep autoencoding models for unsupervised anomaly segmentation in brain MR images”
In: International MICCAI Brainlesion Workshop
DOI: 10.1007/978-3-030-11723-8_16

3.4.3 Contribution: Bayesian Skip-Autoencoders for Unsupervised Hyperintense Anomaly Detection in High Resolution Brain Mri

The previous contribution yielded promising results, but experiments were limited to a narrow anatomical region centered around the axial midline. The application of the very same architectures to axial brain MR slices from entire MRI volumes yielded both a performance degradation in reconstruction quality and in anomaly segmentation. Performance is highly dependent on reconstruction fidelity, as imperfections in the recovery of fine anatomical details lead to false positive residuals. To account for this issue, skip-connections have been integrated into the auto-encoding model. These enable high fidelity reconstructions. Combined with dropout to prevent the AE from learning an identity mapping, considerably higher segmentation performance of WML and brain tumors is observed while reconstruction fidelity improves dramatically. The stochastic nature of the dropout applied to the skip-connections is further used to turn the Skip-AE into a bayesian model, which allows to aggregate MC-reconstructions into a mean residual and to compute a reconstruction uncertainty for every pixel. The results indicate a high level of confidence on the residuals of anomalous pixels, whereas surounding tissue show substantially higher variance.

Contributions | The author of this thesis was responsible for the main idea of introducing skip-connections in AEs and regularizing them with dropout, for the design and implementation of the proposed framework and the evaluation pipeline, planning of the experiments, data preparation and the writing of the manuscript. Benedikt Wiestler contributed with discussions and feedback on the main ideas of the paper, as well as with datasets and their preparation. Shadi Albarqouni contributed with discussions and feedback on the main ideas of the paper, experimental design and evaluation, and was involved in proof-reading. Nassir Navab contributed with discussions and feedback on the main ideas of the paper and was involved in proof-reading.

Copyright statement | This contribution was originally published as: "**Christoph Baur**, Benedikt Wiestler, Shadi Albarqouni, and Nassir Navab. "Bayesian Skip-Autoencoders for Unsupervised Hyperintense Anomaly Detection in High Resolution Brain Mri." *In 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), pp. 1905-1909. IEEE, 2020.*".

Reprint Denied

The reprint of this publication was rejected on open-access platforms. The publication can be found at (`https://ieeexplore.ieee.org/document/9098686`). The details are provided below.

Publication [14]

Authors: Christoph Baur, Benedikt Wiestler, Shadi Albarqouni, Nassir Navab
Title: "Bayesian Skip-Autoencoders for Unsupervised Hyperintense Anomaly Detection in High Resolution Brain Mri"
In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)
DOI: 10.1109/ISBI45749.2020.9098686

3.4.4 Contribution: Scale-Space Autoencoders for Unsupervised Anomaly Segmentation in Brain MRI

While Skip-AEs provide high-quality reconstructions, determining the appropriate configuration of skip-connections and dropout is intricate. As a result, the suppression of anomalies is hard to control. In this contribution, the Laplacian Pyramid is investigated for UAS tasks. The scale-space representation of healthy brain anatomy is modeled with a set of (variational) AEs to achieve high-fidelity, high-resolution reconstructions of input data. As shown in a variety of experiments, compressing and reconstructing different frequency-bands with a multitude of lightweight AEs or VAEs can be solved more effectively than with traditional AEs. Similar to previous work, anomalies are segmented from residuals between input and reconstructions, but here the reconstruction is obtained from the inverse transformation of the reconstructed Laplacian Pyramid. Previous state-of-the-art is notably outperformed in segmenting MS lesions and brain tumors in three different datasets. It is also shown that the multi-scale nature of scale-space can be exploited to further improve anomaly delineation by aggregating residuals from multiple scales. As a result, lesions of varying size and morphology can be segmented with the help of a single, composite model.

Contributions | The author of this thesis was responsible for the main idea of modeling the scale-space representation of healthy brain anatomy with an ensemble of AEs, for the design and implementation of the proposed framework and the evaluation pipeline, planning of the experiments, data preparation and the writing of the manuscript. Benedikt Wiestler contributed with discussions and feedback on the main ideas of the paper, as well as with datasets and their preparation. Shadi Albarqouni contributed with discussions and feedback on the main ideas of the paper, experimental design and evaluation, and was involved in proof-reading. Nassir Navab contributed with discussions and feedback on the main ideas of the paper and was involved in proof-reading.

Copyright statement | This contribution was originally published as: "**Christoph Baur**, Benedikt Wiestler, Shadi Albarqouni, and Nassir Navab. "Scale-Space Autoencoders for Unsupervised Anomaly Segmentation in Brain MRI." *In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 552-561. Springer, Cham, 2020*.".

Reprint Denied

The reprint of this publication was rejected on open-access platforms. The publication can be found at (`https://link.springer.com/chapter/10.1007%2F978-3-030-59719-1_54`). The details are provided below.

Publication [17]

Authors: Christoph Baur, Benedikt Wiestler, Shadi Albarqouni, Nassir Navab
Title: "Scale-Space Autoencoders for Unsupervised Anomaly Segmentation in Brain MRI"
In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings
DOI: https://doi.org/10.1007/978-3-030-59719-1_54

3.4.5 Not peer-reviewed: Autoencoders for Unsupervised Anomaly Segmentation in Brain MR Images: A Comparative Study

This work did not undergo peer-review and is not relevant for the grading of this dissertation.

A more elaborate, full-grown comparative study of AEs for UAS in brain MRI is presented in [11]. This work compares numerous state-of-the-art methods proposed for UAS in medical imaging based on AEs, VAEs and also GANs. A primary goal of this study is to establish comparability among existing methods by training and evaluating on the same datasets at the same image resolution with a single “unified” network architecture for all models. The compared methods are ranked according to different criteria such as segmentation performance and reconstruction quality. Further, correlation effects between these criteria are examined. A variety of other aspects are investigated as well, among which are the number of healthy training data and their influence on UAS performance, the sensitivity of the different methods to domain shift, the effect of regularizing the latent space, different AE bottleneck designs and the performance on different pathologies. Lastly, open research questions and challenges are identified and discussed.

Contributions|The author of this thesis was responsible for the main idea of conducting a comparative study, for its design, the implementation of the proposed the evaluation pipeline, planning of the experiments, data preparation and the writing of the manuscript. Stefan Denner was responsible for the implementation, experimental validation and also contributed to revision of the manuscript. Benedikt Wiestler contributed with discussions and feedback on the main ideas of the paper, as well as with datasets and their preparation. Shadi Albarqouni contributed with discussions and feedback on the main ideas of the paper, experimental design and evaluation, and was involved in proof-reading. Nassir Navab contributed with discussions and feedback on the main ideas of the paper and was involved in proof-reading.

Copyright statement|This work was originally published as: "**Christoph Baur**, Stefan Denner, Benedikt Wiestler, Shadi Albarqouni, and Nassir Navab. "Autoencoders for Unsupervised Anomaly Segmentation in Brain MR Images: A Comparative Study." *arXiv preprint arXiv:2004.03271 (2020)*.".

Autoencoders for Unsupervised Anomaly Segmentation in Brain MR Images: A Comparative Study

Christoph Baur, Stefan Denner, Benedikt Wiestler, Shadi Albarqouni and Nassir Navab

Abstract—**Deep unsupervised representation learning has recently led to new approaches in the field of Unsupervised Anomaly Detection (UAD) in brain MRI. The main principle behind these works is to learn a model of normal anatomy by learning to compress and recover healthy data. This allows to spot abnormal structures from erroneous recoveries of compressed, potentially anomalous samples. The concept is of great interest to the medical image analysis community as it i) relieves from the need of vast amounts of manually segmented training data—a necessity for and pitfall of current supervised Deep Learning—and ii) theoretically allows to detect arbitrary, even rare pathologies which supervised approaches might fail to find. To date, the experimental design of most works hinders a valid comparison, because i) they are evaluated against different datasets and different pathologies, ii) use different image resolutions and iii) different model architectures with varying complexity. The intent of this work is to establish comparability among recent methods by utilizing a single architecture, a single resolution and the same dataset(s). Besides providing a ranking of the methods, we also try to answer questions like i) how many healthy training subjects are needed to model normality and ii) if the reviewed approaches are also sensitive to domain shift. Further, we identify open challenges and provide suggestions for future community efforts and research directions.**

Index Terms—**Anomaly, Segmentation, Detection, Unsupervised, Brain MRI, Autoencoder, Variational, Adversarial, Generative, VAE-GAN, VAEGAN**

I. Introduction

MR imaging of the brain is at the heart of diagnosis and treatment of neurological diseases. When sifting MR scans, Radiologists intuitively rely on a learned model of normal brain anatomy to detect pathologies. However, reading and interpreting MR scans is an intricate process: It is estimated that in 5-10% of scans, a relevant pathology is missed [1]. Recent breakthroughs in machine learning have led to automated medical image analysis methods which achieve great levels of performance in the detection of tumors or lesions arising from neuro-degenerative diseases such as Alzheimers or Multiple Sclerosis (MS). Despite all their outstanding performances, these methods—mainly based on Supervised Deep Learning—carry some disadvantages: 1) their training calls for large and diverse annotated datasets, which are scarce and costly to obtain; 2) the resulting models are limited to the discovery of lesions which are similar to those in the training data. This is especially crucial for rare diseases, for which collecting training data poses a great challenge. Lately, there have been some Deep Learning-driven attempts towards automatic brain pathology detection which tackle the problem from the perspective of so-called Unsupervised Anomaly Detection (UAD). These approaches are more similar to how Radiologists read MR scans, do not require data with pixel-level annotations and have the potential to detect arbitrary anomalies without a-priori knowing about their appearances.

UAD has a long history in medical image analysis and in brain imaging in particular. Traditional methods are based on statistical modeling, content-based retrieval, clustering or outlier-detection. A review on such classical approaches with a focus on brain CT imaging is given in [2]. Since the rise of Deep Learning, a plethora of new, data-driven approaches has appeared. Initially, Autoencoders (AEs), with their ability to learn non-linear transformations of data onto a low-dimensional manifold, have been leveraged for cluster-based anomaly detection. Lately, a variety of works used AEs and generative modeling to not simply detect, but localize and segment anomalies directly in image-space from imperfect reconstructions of input images, which is surveyed in this work in the context of brain MRI.

The underlying idea thereby is to model the distribution of healthy anatomy of the human brain with the help of deep (generative) representation learning. Once trained, anomalies can be detected as outliers from the modeled, normative distribution. AEs [3][4] and their generative siblings [3][5][6][7][8] have emerged as a popular framework to achieve this by essentially learning to compress and reconstruct MR data of healthy anatomy. The respective methods can essentially be divided into two categories: 1) Reconstruction-based approaches compute a pixel-wise discrepancy between input samples and their feed-forward reconstructions to determine anomalous lesions directly in image-space; 2) Restoration-based methods [9][10] try to alter an input image by moving along the latent manifold until a normal counterpart to the input sample is found, which in turn is used again to detect lesions from the pixel-wise discrepancy of the input data and its healthy restoration. To date—albeit all of these methods report promising performances—results can hardly be compared and drawing general conclusions on their strengths & weaknesses is barely possible. This is hindered by the following issues:

C. Baur, S. Denner, S. Albarqouni, N. Navab are with the Chair for Computer Aided Medical Procedures (CAMP), TU Munich, Boltzmannstr. 3, Garching near Munich

S. Albarqouni is with the Computer Vision Laboratory, ETH Zurich, Sternwartstrasse 7, Zurich, Switzerland

B. Wiestler is with the Neuroradiology Department of Klinikum Rechts der Isar, Ismaningerstr. 22, Munich, Germany

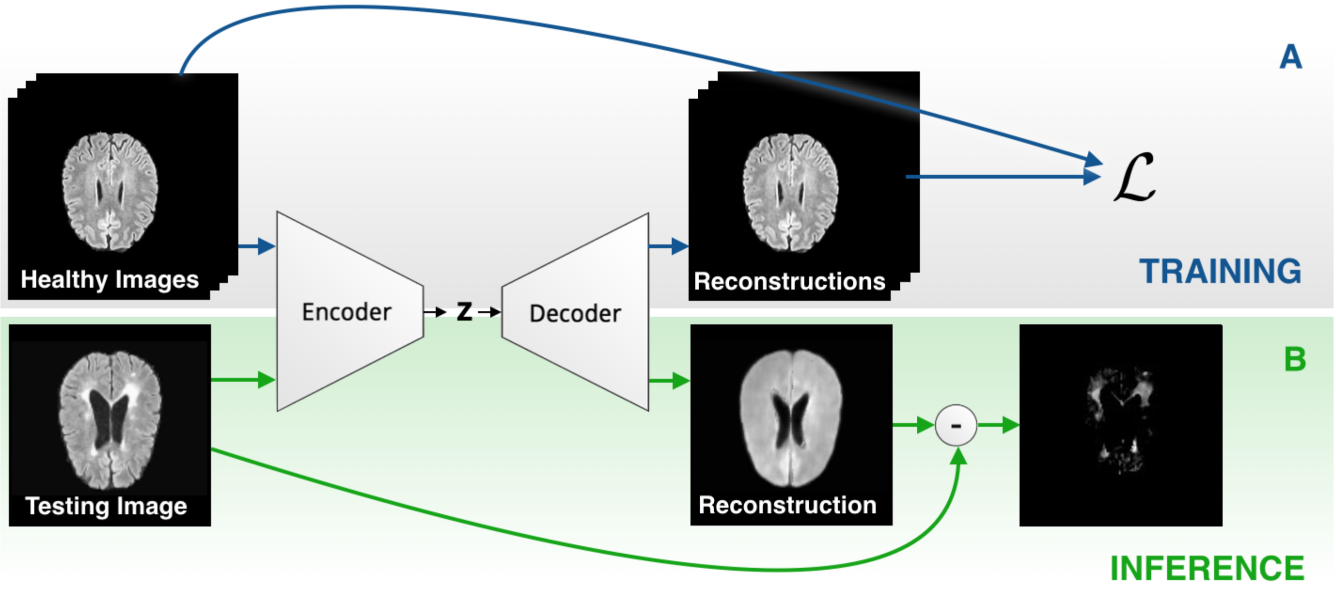


Fig. 1. The concept of Autoencoder-based Anomaly Detection/Segmentation: A) Training a model from only healthy samples and B) anomaly segmentation from erroneous reconstructions of input samples, which might carry an anomaly.

i) most of the works rely on very different datasets with barely overlapping characteristics for their evaluation, ii) are evaluated against different pathologies, iii) operate on different resolutions and iv) utilize different model architectures with varying model complexity. The main intent of this work is to establish comparability among a broad selection of recent methods by utilizing—where applicable—a single network architecture, a single resolution and the same dataset(s).

Contribution—Here, we provide a comparative study of recent Deep-Learning based UAD approaches for brain MRI. We compare various reconstruction- as well as restoration based methods against each other on a variety of different MR datasets with different pathologies[1]. The models are tested on four different datasets for detecting two different pathologies. To evaluate the methods without having to make general assumptions about what constitutes a detection, we utilize pixel-wise segmentation measures as a tight proxy for UAD performance. For a fair comparison, we determined a single, unified architecture on which all the methods rely in this study. This ensures that model complexity is the same for all approaches, if applicable. The performances of the originally proposed networks are also presented. Further, we provide insights on the number of healthy training samples and their impact on model performance, and peek at generalization capabilities of AE models.

II. Unsupervised Deep Representation Learning for Anomaly Detection

A. *Modeling Healthy Anatomy*

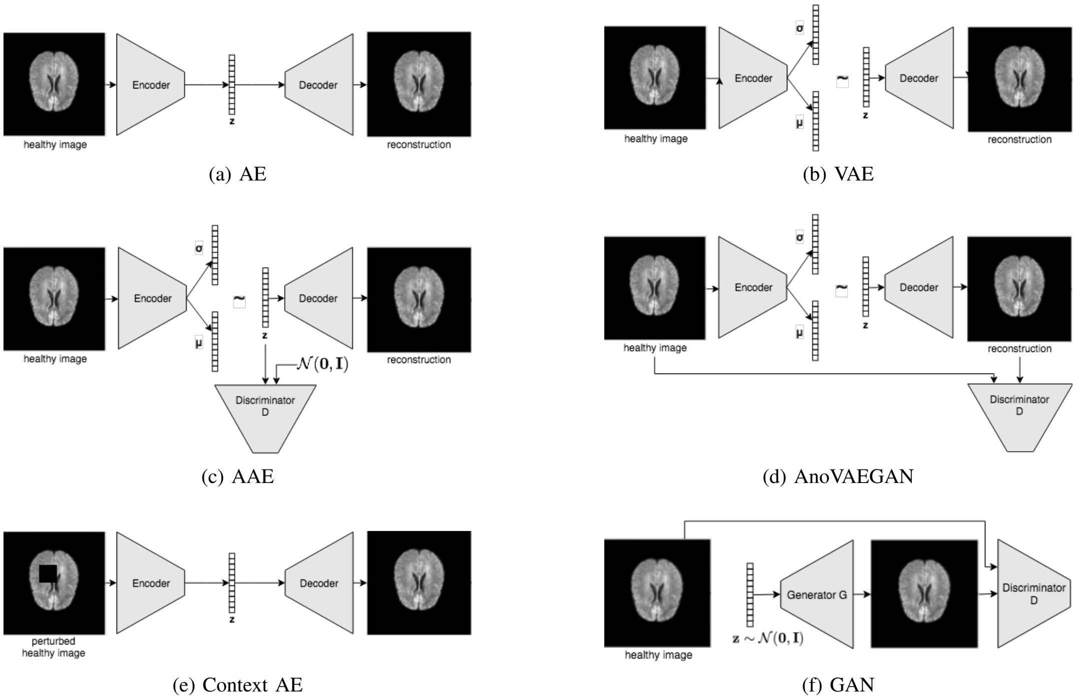
The core concept behind the reviewed methods is the modeling of healthy anatomy with unsupervised deep (generative) representation learning. Therefor, the methods leverage a set of healthy MRI scans $\mathcal{X}_{healthy} \in \mathcal{R}^{D\times H\times W}$ and learn to project it to and recover it from a lower dimensional distribution $\mathbf{z} \in \mathcal{R}^K$ (see Fig. 1). In the following, we first shed the light on the ways how this normative distribution can be modeled, and then present different approaches how anomalies can be discovered using trained models.

Autoencoders—Early work in this field relied on classic AEs (Fig. 2a) to model the normative distribution: An encoder network $\mathrm{Enc}_\theta(\mathbf{x})$ with parameters θ is trained to project a healthy input sample $\mathbf{x} \in \mathcal{X}_{healthy}$ to a lower dimensional manifold $\mathbf{z}$, from which a decoder $\mathrm{Dec}_\phi(\mathbf{z})$ with parameters ϕ then tries to reconstruct the input as $\hat{\mathbf{x}} = \mathrm{Dec}_\phi(\mathrm{Enc}_\theta(\mathbf{x}))$. In other words, the model is trained to compress and reconstruct healthy anatomy by minimizing a reconstruction loss $\mathcal{L}$

$$\arg\min_{\phi,\theta} \mathcal{L}_{AE}^{\phi,\theta}(\mathbf{x},\hat{\mathbf{x}}) = \mathcal{L}_{Rec}^{\phi,\theta}(\mathbf{x},\hat{\mathbf{x}}) = \ell_1(\mathbf{x},\hat{\mathbf{x}}) \tag{1}$$

, which in our case is the ℓ_1-distance between input and reconstruction. The rationale behind this is the assumption that an AE trained on only healthy samples cannot properly reconstruct anomalies in pathological data. This approach has been successfully applied to anomaly segmentation in brain MRI [3][4] and in head CT [11]. A slightly different attempt was made in [7], where the reconstruction-problem was turned into an inpainting-task using a Context Autoencoder (Context AE) (Fig. 2e), in which the model is trained to recover missing sections in healthy training images. The natural choice for the shape of $\mathbf{z}$, here also referred to as *latent space*, *bottleneck* or *manifold*, is a 1D vector. However, it has been shown that spatial AEs with a tensor-shaped bottleneck can be beneficial for high-resolution brain MRI as they preserve spatial context and can generate higher quality reconstructions [3].

[1] Code will be made publicly available at https://github.com/StefanDenn3r/unsupervised_anomaly_detection_brain_mri after successful peer-review of the manuscript.



(a) AE

(b) VAE

(c) AAE

(d) AnoVAEGAN

(e) Context AE

(f) GAN

Fig. 2. Autoencoder-based architectures for UAD at a glance

Latent Variable Models—In classic AEs, there is no regularization on the manifolds structure. In contrast, latent variable models such as Variational Autoencoders (VAEs[12], Fig. 2b) constrain the latent space by leveraging the encoder and decoder networks of AEs to parameterize a latent distribution $q(\mathbf{z}) \sim \mathcal{N}(\mathbf{z}_\mu, \mathbf{z}_\sigma)$, using the following objective:

$$\begin{aligned}\arg\min_{\phi,\theta} \mathcal{L}^{\phi,\theta}_{VAE}(\mathbf{x},\hat{\mathbf{x}}) &= \mathcal{L}^{\phi,\theta}_{Rec}(\mathbf{x},\hat{\mathbf{x}}) + \lambda_{KL}\mathcal{L}^{\theta}_{KL}(q(\mathbf{z}),p(\mathbf{z})) \\ &= \ell_1(\mathbf{x},\hat{\mathbf{x}}) + \lambda_{KL}\mathcal{D}_{KL}(q(\mathbf{z})||p(\mathbf{z}))\end{aligned}$$

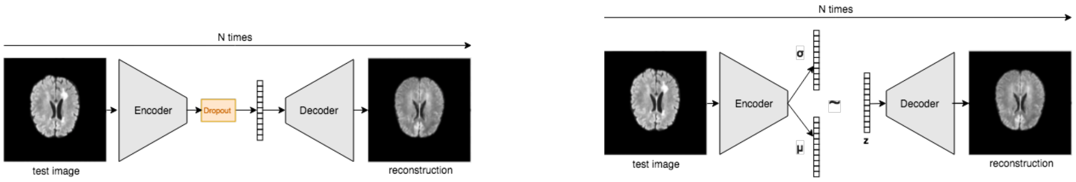
, where λ_{KL} is a Lagrangian multiplier which weights the reconstruction loss against the distribution-matching KL-Divergence $\mathcal{D}_{KL}(\cdot||\cdot)$. In practice, the VAE projects input data onto a learned mean μ and variance σ, from which a sample is drawn and then reconstructed (see Fig. 2b). While the VAE tries to match $q(\mathbf{z})$ to a prior $p(\mathbf{z})$ (typically a multivariate normal distribution) by minimizing the KL-Divergence, which has various shortcomings, the so-called Adversarial Autoencoder (AAE [13], Fig. 2c) leverages an adversarial network as a proxy metric to minimize this discrepancy between the learned distribution $q(\mathbf{z})$ and the prior $p(\mathbf{z})$. As opposed to the KL-Divergence, the optimization via an adversarial network does not favor modes of distributions and is always differentiable. Another extension to the VAE, the so-called Gaussian Mixture VAE (GMVAE [14]) even replaces the mono-modal prior of the VAE with a gaussian mixture, leading to higher expressive power. Due to their ability to model the underlying distribution of high dimensional data, these frameworks are naturally suited for modeling the desired normative distribution. Further, their probabilistic nature facilitates the development of principled density-based anomaly detection methods. Consequently, they have been widely employed for outlier-based anomaly detection: VAEs were used in brain MRI for MS lesion [3], tumor and stroke detection [7]. They have also been utilized for tumor detection in head CT [8] from aggregate means of Monte-Carlo reconstructions. In brain MRI, AAE-[5] and GMVAE[10]-based approaches have also been successfully employed for tumor detection.

Generative Adversarial Networks—Pioneering work, even before AEs were successfully applied for UAD in medical imaging, leveraged Generative Adversarial Networks (GANs [15], Fig. 2f) to detect anomalies in OCT data. Therefor, Schlegl et al [9] modeled the distribution of healthy retinal patches with GANs and determined anomalies by computing the discrepancy between the retinal patch and a healthy counterpart restored by the GAN. Inspired by this work, Baur et al. [3] leveraged the VAEGAN [16]—a combination of the GAN and VAE (Fig. 2d)—to overcome the training instabilities of the GAN and to allow for faster feed-forward inference, which they successfully employed for anomaly segmentation in brain MRI. In recent follow-up work, Schlegl et al. [17] improved on their GAN and also introduced an efficient way to replace the costly iterative restoration method by a single forward pass through the network.

B. Anomaly Segmentation

The trained models can be used for anomaly detection & segmentation in a variety of ways, which are summarized in the following. The interested reader is referred to the original papers for more detailed information.

Reconstruction Based Methods—Such approaches rely on pixel-wise residuals obtained from the difference



(a) Bayesian AE

(b) Bayesian VAE

Fig. 3. Monte Carlo Reconstructions aggregate and average N reconstructions for a single sample.

$$\mathbf{r} = |\mathbf{x} - \hat{\mathbf{x}}| \tag{2}$$

of input samples $\mathbf{x}$ and their reconstruction $\hat{\mathbf{x}}$ (see Fig. 1). The underlying idea being that anomalous structures, which have never been seen during training, cannot be properly reconstructed from the distribution encoded in the latent space, such that reconstruction errors will be high for anomalous structures.

Monte Carlo Methods—For non-deterministic generative models such as VAEs, multiple reconstructions can be obtained by Monte-Carlo (MC) sampling the latent space and an average consensus residual can be computed [8]

$$\mathbf{r} = \frac{1}{N}\sum_{n=1}^{N} |\mathbf{x} - \hat{\mathbf{x}}_{\mathbf{n}}| \tag{3}$$

, with N being the number of MC samplings and $\hat{\mathbf{x}}_{\mathbf{n}}$ being a single MC reconstruction. For deterministic AEs, a similar effect can be achieved by applying dropout with rate p_r to the latent space during inference time, which is also investigated in this work (see Fig. 3a and Fig. 3b for a visual explanation).

Gradient-Based Methods—The gradient-based method proposed in [7] solely relies on image gradients obtained from a single backpropagation step when virtually optimizing for the following objective,

$$\begin{aligned}\arg\min_{\hat{\mathbf{x}}} \mathcal{L}_{Rec}(\mathbf{x}, \hat{\mathbf{x}}) + \lambda_{KL}\mathcal{L}_{KL}(\mathbf{z}, p(\mathbf{z}))& \\ = \ell_1(\mathbf{x}, \hat{\mathbf{x}}) + \lambda_{KL}\mathcal{D}_{KL}(\mathbf{z}||p(\mathbf{z}))& \quad (4)\end{aligned}$$

i.e. the pursuit of bringing the reconstruction $\hat{\mathbf{x}}$ and input $\mathbf{x}$ of the model together while simultaneously moving the latent representation of an input sample closer to the prior (the normal distribution). The resulting pixel-wise gradients are used as a saliency map for anomalies, where it is assumed that stronger gradients constitute anomalies.

Restoration Based Methods—In contrast to reconstruction based methods, restoration based methods involve an optimization on the latent manifold. In the pioneering approach using GANs [9], the goal is to iteratively move along the GANs input distribution $\mathbf{z}$ until a healthy variant of a query image is reconstructed well. Similarly, the method in [10] tries to restore a healthy counterpart $\hat{\mathbf{x}}$ of an input sample $\mathbf{x}$, but by altering it until the ELBO of its latent representation $\mathbf{z}$ is maximized. This can be achieved by initializing $\hat{\mathbf{x}} = \mathbf{x}$ and then iteratively optimizing $\hat{\mathbf{x}}$ for the objective in Eq. 4. Again, the anomalies can be detected in image space from residual maps $\mathbf{r}$ (see Eq. 2).

TABLE I

TRAINING, VALIDATION, TESTING SUBJECTS OF THE DATASETS USED IN THIS STUDY

Dataset	Training	Validation	Testing
$\mathcal{D}_{healthy}$	110	28	-
$\mathcal{D}_{MS}$	-	3	45
$\mathcal{D}_{GB}$	-	-	28
$\mathcal{D}_{MSSEG2015}$	-	-	20
$\mathcal{D}_{MSLUB}$	-	-	30

III. EXPERIMENTS

In the following, we first introduce the datasets used in the experiments, together with their pre-processing, and then introduce the unified network architecture which is the foundation of all the subsequently investigated models. We further explain our post-processing pipeline and all the metrics used in our investigations, before we finally present and discuss the results from various perspectives.

A. Datasets

For this survey, we rely on three different datasets. Selection criteria for these datasets were i) the availability of corresponding T1, T2 and FLAIR scans per subject to be able to leverage a single shared preprocessing pipeline and ii) each dataset being produced with a different MR device.

Healthy, MS & GB—The primary dataset used in this comparative study is a homogenous set of MR scans of both healthy and diseased subjects, produced with a single Philips Achieva 3T MR scanner. It comprises FLAIR, T2- and T1-weighted MR scans of 138 healthy subjects, 48 subjects with MS lesions and 26 subjects with Glioma. All scans have been carefully reviewed and annotated by expert Neuro-Radiologists. Informed consent was waived by the local IRB.

MSLUB—The second MRI dataset [18] consists of co-registered T1, T2 and FLAIR scans of 30 different subjects with MS. Images have been acquired with a 3T Siemens Magnetom Trio MR system at the University Medical Center Ljubljana (UMCL). A gold standard segmentation was obtained from consensus segmentations of three expert raters.

MSSEG2015—The third MRI dataset in our experiments is the publicly available training set of the 2015 Longitudinal MS lesion segmentation challenge [19], which contains 21

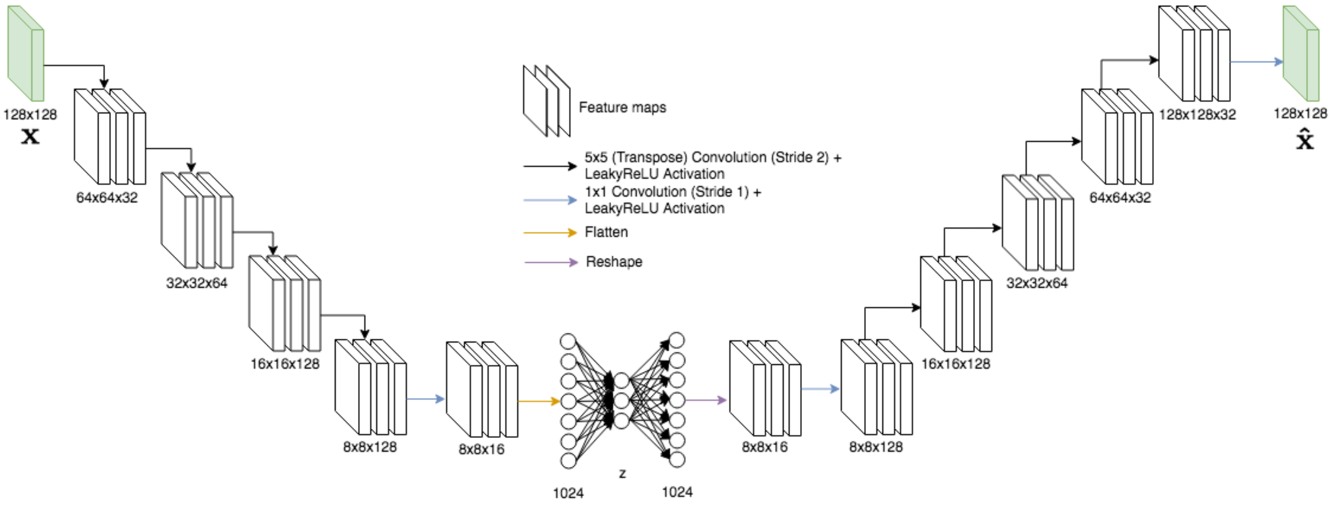


Fig. 4. The unified network architecture with a dense bottleneck. In the case of a spatial bottleneck, the flatten-, dense- and reshape-layers are replaced by a single set of 2D convolutional kernels.

scan sessions from 5 different subjects with T1, T2, PD and FLAIR images each. All data has been acquired with a 3.0 Tesla Philips MRI scanner. The exact device is not known, but the intensity distribution is different from our primary MS & GB datasets. Thus, in this study we utilize the data to test the generalization capabilities of the models and approaches.

Preprocessing and Split—All scans have been brought to the SRI24 ATLAS [20] space to ensure all data share the same volume size and orientation. In succession, the scans have been skull-stripped with ROBEX [21] and denoised with CurvatureFlow [22]. Prior to feeding the data to the networks, all volumes have been normalized into the range [0,1] by dividing each scan by its 98th percentile. All datasets have randomly been split (patient-wise) into training, validation and testing sets as listed in Table I. Training and testing is done on all axial slices of each volume for which the corresponding brainmask indicates the presence of brain pixels. Modus operandi is at a slice resolution of 128 × 128px. This is in stark contrast to some other works, which restrict themselves anatomically to the axial midline [3] or lower resolution [8], [5].

B. Network Architecture and Models

The unified architecture depicted in Fig. 4 was empirically determined in a manual iterative architecture search. The goal was to achieve low reconstruction error on both the training and validation data from $\mathcal{D}_{healthy}$. This unified architecture was then used to train a great variety of models coming from the different, previously introduced domains:

Autoencoders—As a baseline, we used the unified architecture to train a variety of non-generative AEs:

1) **AE (dense)**: an AE with a dense bottleneck $\mathbf{z} \in \mathcal{R}^{128}$
2) **AE (spatial)** [3]: an AE with a spatial bottleneck $\mathbf{z} \in \mathcal{R}^{8\times8\times128}$
3) **Context AE** [7]: with $\mathbf{z} \in \mathcal{R}^{128}$
4) **Constrained AE** [5]: with $\mathbf{z} \in \mathcal{R}^{128}$

Latent Variable Models—Further, we trained various generative latent variable models using the same unified architecture and bottleneck configurations:

1) **VAE** [3], [6]: with $\mathbf{z} \in \mathcal{R}^{128}$
2) **Context VAE** [7]: with $\mathbf{z} \in \mathcal{R}^{128}$
3) **Constrained AAE** [5]: with $\mathbf{z} \in \mathcal{R}^{128}$
4) **GMVAE (dense)** [10]: with $\mathbf{z} \in \mathcal{R}^{128}$
5) **GMVAE (spatial)** [10]: with $\mathbf{z} \in \mathcal{R}^{8\times8\times128}$

Generative Adversarial Networks—Finally, we also trained an AnoVAEGAN [3] and an f-AnoGAN [17], whose encoder-decoder networks implement the unified architecture, and the discriminator network is a replica of the encoder:

1) **AnoVAEGAN** [3]: with $\mathbf{z} \in \mathcal{R}^{128}$
2) **fAnoGAN** [17]: with $\mathbf{z} \in \mathcal{R}^{128}$

Noteworthy, both methods were optimized with the Wasserstein loss [23] to avoid GAN training instabilities and mode collapse.

All models were trained from $\mathcal{D}_{healthy}$ until convergence using an automatic early stopping criterion, i.e. training was stopped if the reconstruction loss on the held-out validation set from $\mathcal{D}_{healthy}$ did not improve more than an $\epsilon > 10e-9$ for 5 epochs. In succession, all the methods were used for reconstruction-based anomaly detection. The trained VAE and GMVAE were also used for the density-based image restoration [10], where each sample was restored in 500 iterations:

1) **VAE (restoration)** [10]
2) **GMVAE (restoration)** [10]

Both AE (dense) and VAE were also used for MC-reconstruction based anomaly detection:

1) **Bayesian AE** [8]: Dropout rate 0.2
2) **Bayesian VAE** [8]: $N = 100$ MC-samples per input slice

and in the case of the Context VAE, we also tried the gradient-based approach proposed in [7]:

TABLE II
HYPERPARAMETERS FOR THE DIFFERENT MODELS

Param	**Value**
learning rate	0.0001
λ_{KL}	1.0
dropout rate p_r	0.2

1) **Context VAE (gradient)** [7]

Hyperparameters can be taken from Table II.

C. Postprocessing

The output of all models and approaches is subject to the same post-processing. Every residual image $\mathbf{r}$ is first multiplied with a slightly eroded brain-mask to remove prominent residuals occuring near sharp edges at brain-mask boundaries and gyri and sulci (the latter are very diverse and hard to model). Further, for the MS lesion datasets we make use of prior knowledge and only keep positive residuals as these lesions are known to be fully hyper-intense in FLAIR images. For each MR volume, the residual images for all slices are first aggregated into a corresponding 3D residual volume, which is then subject to a 3D median filtering with a $5 \times 5 \times 5$ kernel to remove small outliers and to obtain a more continuous signal. The latter is beneficial for the subsequent model assessment as it leads to smoother curves. As a final step, the continuous output is binarized and a 3D connected component analysis is performed on the resulting binary volumes to discard any small structures with an area less than 8 voxels.

D. Metrics

We assess the anomaly segmentation performance at a level of single voxels, at which class imbalance needs careful consideration as anomalous voxels are usually less frequent than normal voxels. To do so, we generate dataset-specific Precision-Recall-Curves (PRC) and then compute the area under it (AUPRC). Noteworthy, this allows to judge the models capabilities without choosing an Operating Point (OP). Further, for each model we provide an estimate of its theoretically best possible DICE-score ($\lceil$DICE$\rceil$) on each dataset. Therefor, for each testing dataset $d \in \mathcal{D} = \mathcal{D}_{MS}, \mathcal{D}_{GB}, \mathcal{D}_{MSSEG2015}, \mathcal{D}_{MSLUB}$, we utilize the available ground-truth segmentation and perform a greedy search up to three decimals to determine the respective OP on the PRC curve which yields the best possible DICE score for dataset d. Additionally, to simulate the models performance in more realistic settings, we utilize a held-out validation set from $\mathcal{D}_{MS}$ to determine an OP t at which we then compute patient-specific DICE-scores for every dataset. Some of the reviewed works originally utilize Receiver-Operating-Characteristics (ROC) to evaluate anomaly detection performance. We report the area under such ROC curves (AUROC) as well, but want to emphasize that it has to be used with care. Under heavy class imbalance, ROC curves can be misleading as they give much higher weight to the more frequent class and thus in the case of a pixel-wise assessment very optimistic views on performance.

To gain deeper insights what makes a model capable of segmenting anomalies better than others, we also report dataset-specific ℓ_1-reconstruction errors on normal (ℓ_1-RE$_N$) and anomalous voxels (ℓ_1-RE$_A$), as well as the $\mathcal{X}^2$-distance of the respective normal and anomalous residual histograms for every model.

E. Overview

Detailed results of all models and UAD approaches on all datasets can be found in Tables III ($\mathcal{D}_{MS}$), IV ($\mathcal{D}_{GB}$), V ($\mathcal{D}_{MSLUB}$) and VI ($\mathcal{D}_{MSSEG2015}$). In the following, we analyze all these data from different perspectives. We start by first comparing different model types and bottleneck design, followed by the different ways to detect anomalies directly in image-space. Then, we shed the light on the number of training subjects and their impact on performance, and elaborate on domain shift.

F. Constraining & Regularization

Initially, we compare the classic AE (dense) to its VAE and Constrained AE counterpart to investigate the effect of constraining or regularizing the latent space of the models. Recall that VAEs regularize the latent space to follow a prior distribution, whereas the deterministic Constrained AE enforces that reconstructions and input lie closely on the manifold. We measure the models' performances in terms of the AUPRC as well as the $\lceil$DICE$\rceil$ and glimpse at the reconstruction errors for normal and anomalous pixels (see Fig. 11 for residual histograms of normal and anomalous voxels). We see from Table III that explicitly modeling a distribution with a VAE leads to dramatic performance gains on $\mathcal{D}_{MS}$ over the standard AE, and introducing the matching constraint (Constrained AE) between $\mathbf{x}$ and $\hat{\mathbf{x}}$ improves the performance even more. On all other datasets, the VAE clearly is the winner among the compared models, but the Constrained AE still outperforms the classic AE. From these results, we deduce that enforcing a structure on the manifold of AEs is indeed beneficial for UAD.

G. Dense vs Spatial Bottleneck

To determine if the design of the AE bottleneck can improve the performance of the models, we further compare dense models for which a spatial counterpart exists, i.e. AE (dense) vs AE (spatial) vs GMVAE (dense) vs GMVAE (spatial). The spatial bottleneck allows the model to preserve spatial information and geometric features in its latent space, which positively affects the models reconstruction capabilities. From Tables III-VI it can be seen that the dense models outperform the spatial variants, alone the spatial AE performs slightly better on $\mathcal{D}_{GB}$ than its dense counterpart. We find that at our resolution of 128x128px, the spatial models reconstruct their input too well (see Fig. 7), including the anomalies.

H. Latent Variable Models

Next, we focus only on different latent variable model types, i.e. the VAE, GMVAE (dense) and Constrained AAE. On the

TABLE III
EXPERIMENTAL RESULTS ON THE MS DATASET

Approach	**AUROC**	**AUPRC**	$\lceil$**DICE**$\rceil$	**DICE** $(\mu \pm \sigma)$	ℓ_1-**RE**$_N$ $(\mu \pm \sigma)$	ℓ_1-**RE**$_A$ $(\mu \pm \sigma)$	$\mathcal{X}^2$
AE (dense)	0.918	0.271	0.389	0.325 ± 0.164	5.30e-10 ± 5.45e-06	4.07e-08 ± 4.29e-05	3.59e-01
AE (spatial) [3]	0.852	0.13	0.231	0.165 ± 0.134	3.86e-10 ± 1.91e-06	1.78e-08 ± 1.25e-05	8.39e-02
VAE [3], [6]	0.945	0.399	0.469	0.389 ± 0.166	8.27e-10 ± 8.11e-06	7.30e-08 ± 6.32e-05	4.46e-01
VAE (restoration) [10]	0.946	0.454	0.495	0.404 ± 0.176	8.92e-10 ± 8.49e-06	7.77e-08 ± 6.63e-05	4.61e-01
Context AE [7]	0.9	0.233	0.374	0.327 ± 0.173	6.27e-10 ± 6.78e-06	6.68e-08 ± 6.01e-05	4.26e-01
Context VAE [7]	0.937	0.416	0.492	0.418 ± 0.167	6.09e-10 ± 6.27e-06	7.15e-08 ± 5.93e-05	4.17e-01
Context VAE (gradient) [7]	0.963	0.294	0.385	0.305 ± 0.161	9.00e-10 ± 8.38e-06	3.18e-08 ± 4.68e-05	2.77e-01
GMVAE (dense) [10]	0.944	0.389	0.477	0.387 ± 0.178	8.11e-10 ± 8.08e-06	7.60e-08 ± 6.44e-05	4.51e-01
GMVAE (dense restoration) [10]	0.945	0.453	0.501	0.411 ± 0.180	9.07e-10 ± 8.64e-06	8.34e-08 ± 6.84e-05	4.77e-01
GMVAE (spatial) [10]	0.877	0.096	0.191	0.148 ± 0.126	3.98e-10 ± 1.91e-06	1.57e-08 ± 1.02e-05	7.91e-02
GMVAE (spatial restoration) [10]	0.925	0.295	0.363	0.287 ± 0.157	2.81e-10 ± 2.04e-06	1.60e-08 ± 1.31e-05	1.12e-01
f-AnoGAN [17]	0.957	0.448	0.489	0.417 ± 0.178	2.05e-09 ± 1.54e-05	1.06e-07 ± 7.28e-05	4.96e-01
AnoVAEGAN [3]	0.947	0.376	0.45	0.371 ± 0.178	1.22e-09 ± 1.10e-05	9.78e-08 ± 7.37e-05	5.03e-01
Constrained AE [5]	0.94	0.429	0.485	0.409 ± 0.173	5.73e-10 ± 5.68e-06	4.16e-08 ± 4.29e-05	3.64e-01
Constrained AAE [5]	0.949	0.268	0.392	0.331 ± 0.195	1.66e-09 ± 1.42e-05	1.04e-07 ± 7.94e-05	5.06e-01
Bayesian AE [8]	0.913	0.262	0.373	0.313 ± 0.159	5.44e-10 ± 5.56e-06	4.23e-08 ± 4.36e-05	3.72e-01
Bayesian VAE [8]	0.945	0.403	0.471	0.390 ± 0.165	8.23e-10 ± 8.09e-06	7.31e-08 ± 6.37e-05	4.46e-01

TABLE IV
EXPERIMENTAL RESULTS ON THE GB DATASET

Approach	**AUROC**	**AUPRC**	$\lceil$**DICE**$\rceil$	**DICE** $(\mu \pm \sigma)$	ℓ_1-**RE**$_N$ $(\mu \pm \sigma)$	ℓ_1-**RE**$_A$ $(\mu \pm \sigma)$	$\mathcal{X}^2$
AE (dense)	0.753	0.158	0.299	0.268 ± 0.133	1.73e-09 ± 1.14e-05	6.05e-08 ± 6.57e-05	4.20e-01
AE (spatial) [3]	0.737	0.179	0.295	0.239 ± 0.127	7.82e-10 ± 2.78e-06	2.88e-08 ± 1.80e-05	4.08e-02
VAE [3], [6]	0.795	0.272	0.441	0.374 ± 0.162	3.62e-09 ± 2.17e-05	1.52e-07 ± 1.34e-04	6.18e-01
VAE (restoration) [10]	0.8	0.441	0.537	0.435 ± 0.193	2.96e-09 ± 1.73e-05	1.97e-07 ± 1.55e-04	6.45e-01
Context AE [7]	0.753	0.253	0.402	0.343 ± 0.160	1.69e-09 ± 1.18e-05	1.23e-07 ± 9.80e-05	4.28e-01
Context VAE [7]	0.775	0.215	0.375	0.333 ± 0.139	2.27e-09 ± 1.48e-05	1.03e-07 ± 9.20e-05	5.09e-01
Context VAE (gradient) [7]	0.799	0.172	0.315	0.281 ± 0.122	2.85e-09 ± 1.76e-05	5.03e-08 ± 8.02e-05	3.96e-01
GMVAE (dense) [10]	0.798	0.367	0.492	0.406 ± 0.176	2.98e-09 ± 1.79e-05	1.62e-07 ± 1.38e-04	6.21e-01
GMVAE (dense restoration) [10]	0.797	0.423	0.522	0.421 ± 0.190	2.95e-09 ± 1.74e-05	2.04e-07 ± 1.57e-04	6.48e-01
GMVAE (spatial) [10]	0.737	0.119	0.258	0.216 ± 0.125	8.39e-10 ± 3.00e-06	2.19e-08 ± 1.43e-05	5.39e-02
GMVAE (spatial restoration) [10]	0.752	0.21	0.313	0.272 ± 0.128	6.73e-10 ± 3.13e-06	2.06e-08 ± 1.89e-05	9.84e-02
f-AnoGAN [17]	0.786	0.349	0.447	0.379 ± 0.174	5.20e-09 ± 2.54e-05	2.32e-07 ± 1.78e-04	6.92e-01
AnoVAEGAN [3]	0.774	0.334	0.485	0.385 ± 0.191	3.65e-09 ± 2.21e-05	2.33e-07 ± 1.75e-04	6.77e-01
Constrained AE [5]	0.772	0.23	0.353	0.318 ± 0.145	1.80e-09 ± 1.14e-05	7.02e-08 ± 7.35e-05	4.69e-01
Constrained AAE [5]	0.793	0.365	0.481	0.392 ± 0.183	4.12e-09 ± 2.24e-05	2.33e-07 ± 1.75e-04	6.86e-01
Bayesian AE [8]	0.747	0.143	0.28	0.253 ± 0.124	1.77e-09 ± 1.16e-05	5.81e-08 ± 6.42e-05	4.15e-01
Bayesian VAE [8]	0.795	0.271	0.44	0.374 ± 0.162	3.70e-09 ± 2.21e-05	1.53e-07 ± 1.35e-04	6.18e-01

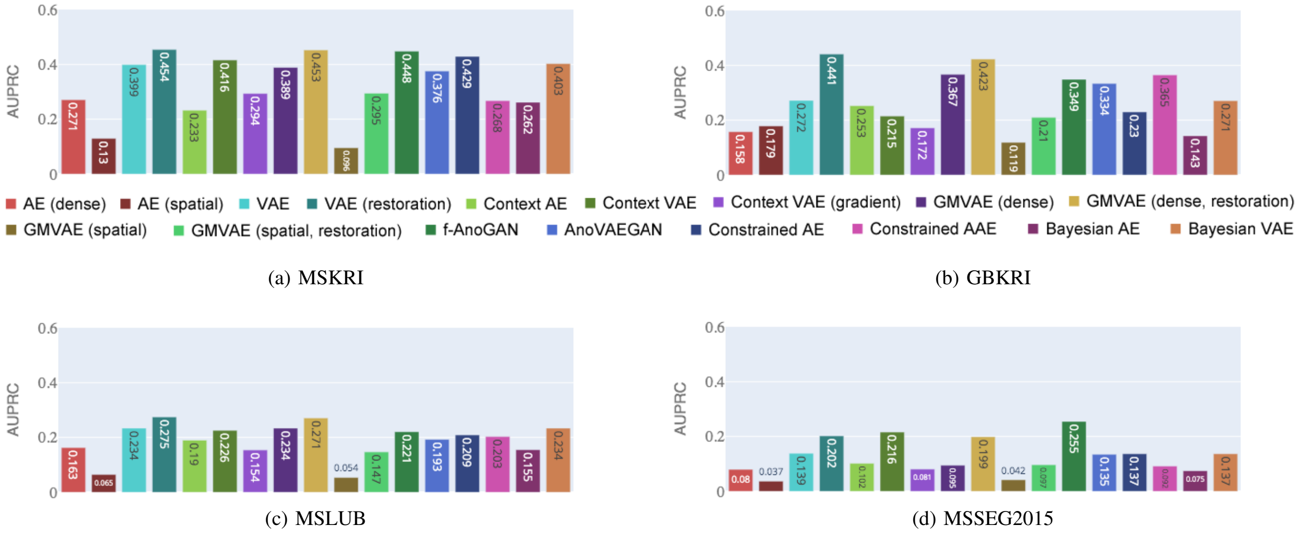
TABLE V
EXPERIMENTAL RESULTS ON THE MSLUB DATASET

Approach	**AUROC**	**AUPRC**	$\lceil$**DICE**$\rceil$	**DICE** $(\mu \pm \sigma)$	ℓ_1-**RE**$_N$ $(\mu \pm \sigma)$	ℓ_1-**RE**$_A$ $(\mu \pm \sigma)$	$\mathcal{X}^2$
AE (dense)	0.794	0.163	0.271	0.181 ± 0.168	9.21e-10 ± 7.01e-06	4.58e-08 ± 5.42e-05	4.35e-01
AE (spatial) [3]	0.732	0.065	0.154	0.098 ± 0.116	7.05e-10 ± 2.58e-06	2.21e-08 ± 1.51e-05	1.04e-01
VAE [3], [6]	0.827	0.234	0.323	0.205 ± 0.207	1.67e-09 ± 1.15e-05	8.36e-08 ± 8.84e-05	5.53e-01
VAE (restoration) [10]	0.839	0.275	0.333	0.203 ± 0.209	1.92e-09 ± 1.27e-05	9.31e-08 ± 9.53e-05	5.65e-01
Context AE [7]	0.771	0.19	0.28	0.193 ± 0.186	9.65e-10 ± 7.32e-06	6.10e-08 ± 6.87e-05	4.90e-01
Context VAE [7]	0.805	0.226	0.316	0.204 ± 0.202	1.12e-09 ± 8.34e-06	6.75e-08 ± 7.27e-05	5.12e-01
Context VAE (gradient) [7]	0.889	0.154	0.265	0.175 ± 0.173	1.56e-09 ± 1.13e-05	4.21e-08 ± 6.06e-05	3.71e-01
GMVAE (dense) [10]	0.832	0.234	0.316	0.202 ± 0.210	1.80e-09 ± 1.22e-05	8.65e-08 ± 8.94e-05	5.56e-01
GMVAE (dense restoration) [10]	0.836	0.271	0.332	0.204 ± 0.208	2.01e-09 ± 1.32e-05	9.87e-08 ± 9.74e-05	5.75e-01
GMVAE (spatial) [10]	0.756	0.054	0.136	0.102 ± 0.106	7.07e-10 ± 2.62e-06	1.95e-08 ± 1.31e-05	1.03e-01
GMVAE (spatial restoration) [10]	0.804	0.147	0.23	0.158 ± 0.149	4.89e-10 ± 2.68e-06	1.81e-08 ± 1.63e-05	1.28e-01
f-AnoGAN [17]	0.856	0.221	0.283	0.189 ± 0.192	4.56e-09 ± 2.39e-05	1.51e-07 ± 1.21e-04	6.26e-01
AnoVAEGAN [3]	0.823	0.193	0.282	0.180 ± 0.167	2.04e-09 ± 1.36e-05	1.01e-07 ± 9.56e-05	5.89e-01
Constrained AE [5]	0.821	0.209	0.298	0.197 ± 0.187	1.11e-09 ± 8.07e-06	4.73e-08 ± 5.61e-05	4.73e-01
Constrained AAE [5]	0.852	0.203	0.289	0.194 ± 0.207	3.35e-09 ± 1.97e-05	1.26e-07 ± 1.12e-04	5.97e-01
Bayesian AE [8]	0.79	0.155	0.267	0.183 ± 0.162	9.39e-10 ± 7.17e-06	4.80e-08 ± 5.57e-05	4.43e-01
Bayesian VAE [8]	0.827	0.234	0.322	0.201 ± 0.206	1.68e-09 ± 1.16e-05	8.36e-08 ± 8.84e-05	5.53e-01

MS datasets $\mathcal{D}_{MS}, \mathcal{D}_{MSSEG2015}$ and $\mathcal{D}_{MSLUB}$, the VAE constitutes the best among the compared models. The Constrained AAE yields lower performance than the other models—also lower than its non-generative sibling, the Constrained AE. However, on the Glioblastoma dataset, it is on par with the GMVAE, and both models significantly outperform the VAE in the detection of brain tumors. Generally, the performance of the GMVAE generally seems to heavily depend on the dataset

TABLE VI
EXPERIMENTAL RESULTS ON THE MSSEG2015 DATASET

Approach	**AUROC**	**AUPRC**	**⌈DICE⌉**	**DICE** $(\mu \pm \sigma)$	ℓ_1-**RE**$_N$ $(\mu \pm \sigma)$	ℓ_1-**RE**$_A$ $(\mu \pm \sigma)$	$\mathcal{X}^2$
AE (dense)	0.879	0.08	0.185	0.150 ± 0.075	1.87e-09 ± 1.10e-05	6.40e-08 ± 5.41e-05	4.90e-01
AE (spatial) [3]	0.781	0.037	0.106	0.066 ± 0.073	1.11e-09 ± 3.34e-06	2.68e-08 ± 1.52e-05	1.40e-01
VAE [3], [6]	0.899	0.139	0.257	0.200 ± 0.124	2.89e-09 ± 1.52e-05	1.10e-07 ± 7.43e-05	6.07e-01
VAE (restoration) [10]	0.905	0.202	0.272	0.211 ± 0.122	3.44e-09 ± 1.69e-05	1.18e-07 ± 7.91e-05	6.10e-01
Context AE [7]	0.877	0.102	0.225	0.188 ± 0.116	1.93e-09 ± 1.16e-05	8.92e-08 ± 6.34e-05	5.49e-01
Context VAE [7]	0.896	0.216	0.336	0.267 ± 0.112	1.74e-09 ± 1.00e-05	9.24e-08 ± 6.47e-05	5.28e-01
Context VAE (gradient) [7]	0.923	0.081	0.173	0.127 ± 0.088	2.39e-09 ± 1.48e-05	6.07e-08 ± 5.85e-05	3.87e-01
GMVAE (dense) [10]	0.9	0.095	0.21	0.174 ± 0.121	2.99e-09 ± 1.64e-05	1.06e-07 ± 7.13e-05	6.05e-01
GMVAE (dense restoration) [10]	0.909	0.199	0.28	0.223 ± 0.124	3.98e-09 ± 1.86e-05	1.31e-07 ± 8.01e-05	6.35e-01
GMVAE (spatial) [10]	0.846	0.042	0.106	0.069 ± 0.073	1.10e-09 ± 3.34e-06	2.73e-08 ± 1.36e-05	1.29e-01
GMVAE (spatial restoration) [10]	0.873	0.097	0.178	0.118 ± 0.110	8.13e-10 ± 3.52e-06	2.51e-08 ± 1.63e-05	1.57e-01
f-AnoGAN [17]	0.923	0.255	0.342	0.278 ± 0.140	3.43e-08 ± 7.40e-05	7.85e-07 ± 1.98e-04	9.48e-01
AnoVAEGAN [3]	0.911	0.135	0.235	0.200 ± 0.133	5.97e-09 ± 2.55e-05	1.91e-07 ± 1.03e-04	6.93e-01
Constrained AE [5]	0.9	0.137	0.261	0.209 ± 0.100	2.26e-09 ± 1.22e-05	6.76e-08 ± 5.33e-05	5.16e-01
Constrained AAE [5]	0.917	0.092	0.204	0.190 ± 0.170	1.14e-08 ± 3.98e-05	2.94e-07 ± 1.24e-04	7.50e-01
Bayesian AE [8]	0.877	0.075	0.176	0.142 ± 0.072	1.89e-09 ± 1.13e-05	6.71e-08 ± 5.46e-05	4.98e-01
Bayesian VAE [8]	0.898	0.137	0.252	0.194 ± 0.117	2.87e-09 ± 1.51e-05	1.08e-07 ± 7.42e-05	6.07e-01



(a) MSKRI

(b) GBKRI

(c) MSLUB

(d) MSSEG2015

Fig. 5. AUPRC of all models and UAD approaches, using the unified architecture.

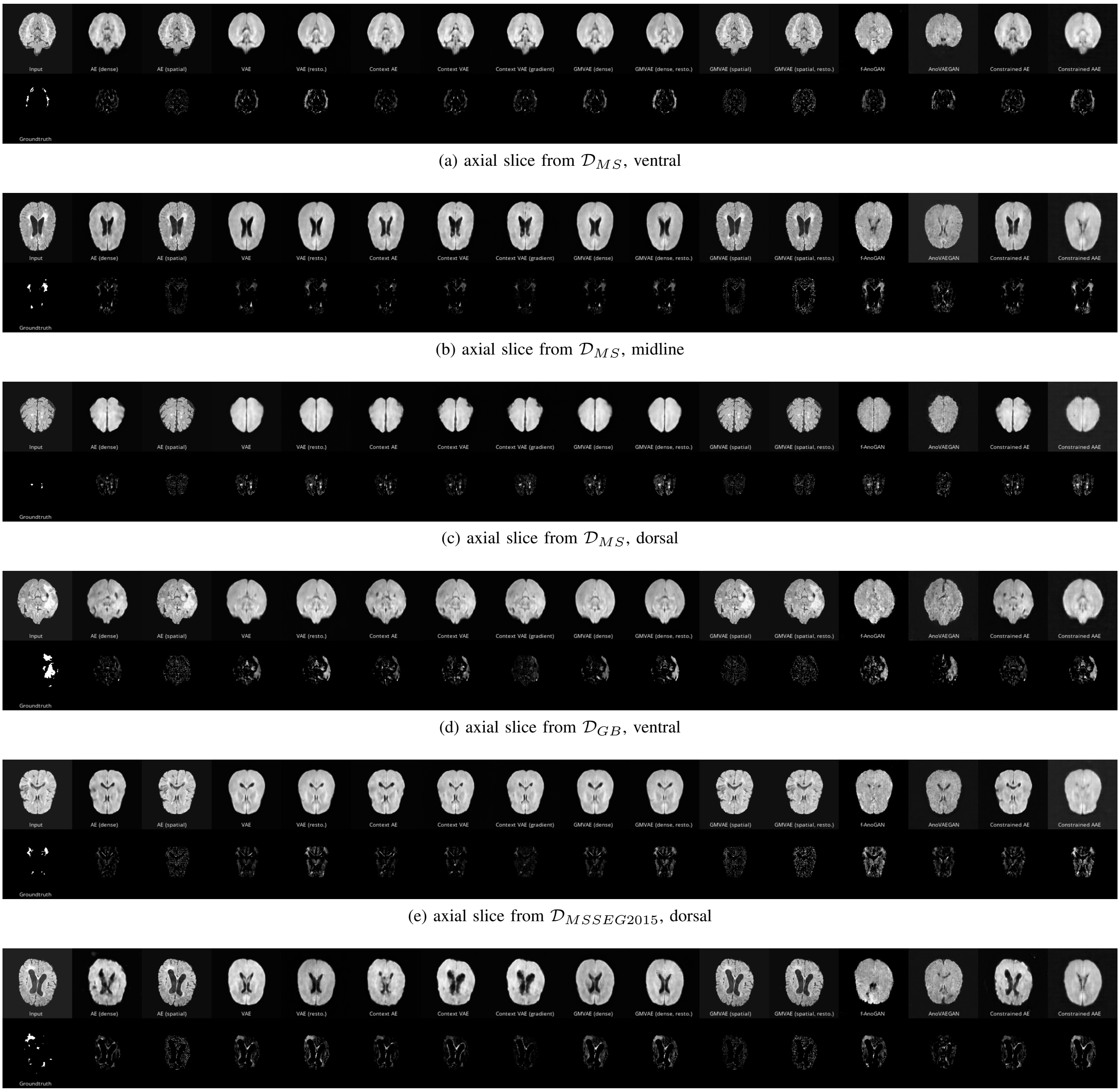
rather than the pathology: On $\mathcal{D}_{MS}$ and $\mathcal{D}_{MSLUB}$ it behaves very similar to the VAE, whereas on $\mathcal{D}_{GB}$ and $\mathcal{D}_{MSSEG2015}$ its performance resembles that of the Constrained AAE.

I. GAN-based models

GAN-based models are known to produce very realistic and crisp images, while AEs are known for their blurry reconstructions. Indeed, qualitative comparison of the f-AnoGAN and the AnoVAEGAN to the AE and VAE shows that the GAN-based models promote sharpness. This is particularly evident near the boundaries of the brain (see Fig. 6). However, both the f-AnoGAN and AnoVAEGAN model the training distribution too well, such that reconstructions often differ anatomically from the actual input samples (see Fig. 6b for an axial midline slice from $\mathcal{D}_{MS}$). This is especially the case for the AnoVAEGAN, which produces the most crisp reconstructions, but often does not preserve anatomical coherence at all. As a result, on the MS datasets its performance is only comparable to the VAE, but it works considerably better for Glioblastoma segmentation. The f-AnoGAN does not provide as crisp images, but preserves the shape of the input sample and the difference between reconstruction residuals on normal and anomalous pixels is considerably higher across all datasets than for any of the other methods. This makes the UAD performance of the f-AnoGAN stand out. In total, both GAN-based approaches significantly outperform the standard AE (on average, more than 9% for the AnoVAEGAN and more than 15% for the f-AnoGAN) and the f-AnoGAN clearly also outperforms the VAE (on average more than 6%).

J. Monte-Carlo Methods

Monte-Carlo methods applied to (variational) AEs provide an interesting means to aggregate a consensus reconstruction, in which only very likely image features should be emphasized. To investigate if anomalies are affected, we experiment with $N = 100$ MC-reconstructions and—where necessary—an empirically chosen dropout-rate $p_d = 0.2$ to trade-off reconstruction quality and chance. We find that, compared



(a) axial slice from $\mathcal{D}_{MS}$, ventral

(b) axial slice from $\mathcal{D}_{MS}$, midline

(c) axial slice from $\mathcal{D}_{MS}$, dorsal

(d) axial slice from $\mathcal{D}_{GB}$, ventral

(e) axial slice from $\mathcal{D}_{MSSEG2015}$, dorsal

(f) axial slice from $\mathcal{D}_{MSLUB}$, midline

Fig. 6. Visual examples of the different reviewed methods on different datasets, using the unified architecture. Top row: reconstructions; Bottom row: raw residuals.

to one-shot reconstructions, the impact of MC-sampling is at most subtle, and not consistent across different models and datasets. A comparison of AE (dense) to the Bayesian AE shows that MC-dropout leads to a slightly worse performance in almost all metrics across all datasets. On the other hand, the Bayesian VAE, which does not need dropout for MC sampling due to its probabilistic bottleneck, is equal to or slightly outperforms the VAE on $\mathcal{D}_{MS}$, but not on $\mathcal{D}_{GB}$ and $\mathcal{D}_{MSLUB}$. Overall, these numbers indicate that MC methods, albeit an interesting approach, do not provide significant gains in the way they are currently employed.

K. Reconstruction vs Restoration

Previous comparisons focused on different model types and all relied on the reconstruction-based UAD concept. In the following, we rank reconstruction-based methods, against gradient- and restoration-based UAD approaches. More precisely, we compare reconstruction against restoration on the VAE, GMVAE (dense) and GMVAE (spatial). We further rank the restoration-based methods against the top-candidate f-AnoGAN. From Tables III to VI it is evident that restoration based UAD is generally superior to the reconstruction-based counterparts (ranging from 4-17% for the VAE, 4-10% for the

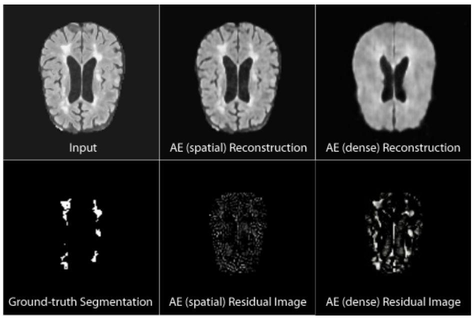

Input
AE (spatial) Reconstruction
AE (dense) Reconstruction
Ground-truth Segmentation
AE (spatial) Residual Image
AE (dense) Residual Image

Fig. 7. Reconstructions and postprocessed residuals using dense and spatial AEs

dense GMVAE and 6-20% for the spatial GMVAE). Consistent with our previously measured results on dense versus spatial models, we also witness a dramatic drop in performance when using the spatial GMVAE, though. Except for $\mathcal{D}_{MSSEG2015}$, the dense restoration methods outperform the f-AnoGAN in all scenarios in terms of the AUPRC and $\lceil$DICE$\rceil$.

L. Domain Shift

Deep Learning models trained from data coming from one domain generally have difficulties to generalize well to other domains, and tackling such domain shift is still a highly active research area. Here, we want to determine to which extent AEs are prone to this effect and if some methods generalize better than others. Subject to our investigations are the MS datasets $\mathcal{D}_{MS}$, $\mathcal{D}_{MSLUB}$ and $\mathcal{D}_{MSSEG2015}$ among which such shifts occur. Generally, UAD performance is best on $\mathcal{D}_{MS}$, which matches the training data distribution, and on both $\mathcal{D}_{MSLUB}$ & $\mathcal{D}_{MSSEG2015}$, the UAD performance drops significantly. However, the reasons for this drop can be manifold, and we want to emphasize that UAD performance as such is not a good indicator for domain shift, as the lesion size and count differs across datasets, and the contrast for $\mathcal{D}_{MS}$ is considerably better than for the other datasets. Instead, we suggest to look at the reconstruction error of normal pixels ℓ_1-RE$_N$ in these datasets. From Tables III, V and VI it can be seen that this error hardly degrades across all these datasets. This implies that generalization measured in terms of the models reconstruction capabilities is not of primary concern. However, from aforementioned tables it can be seen that the reconstruction error of anomalous pixels ℓ_1-RE$_N$ is significantly smaller on $\mathcal{D}_{MSLUB}$ and $\mathcal{D}_{MSSEG2015}$, which is a clear indicator of weaker contrast between normal tissue and lesions in these datasets.

M. Different Pathologies

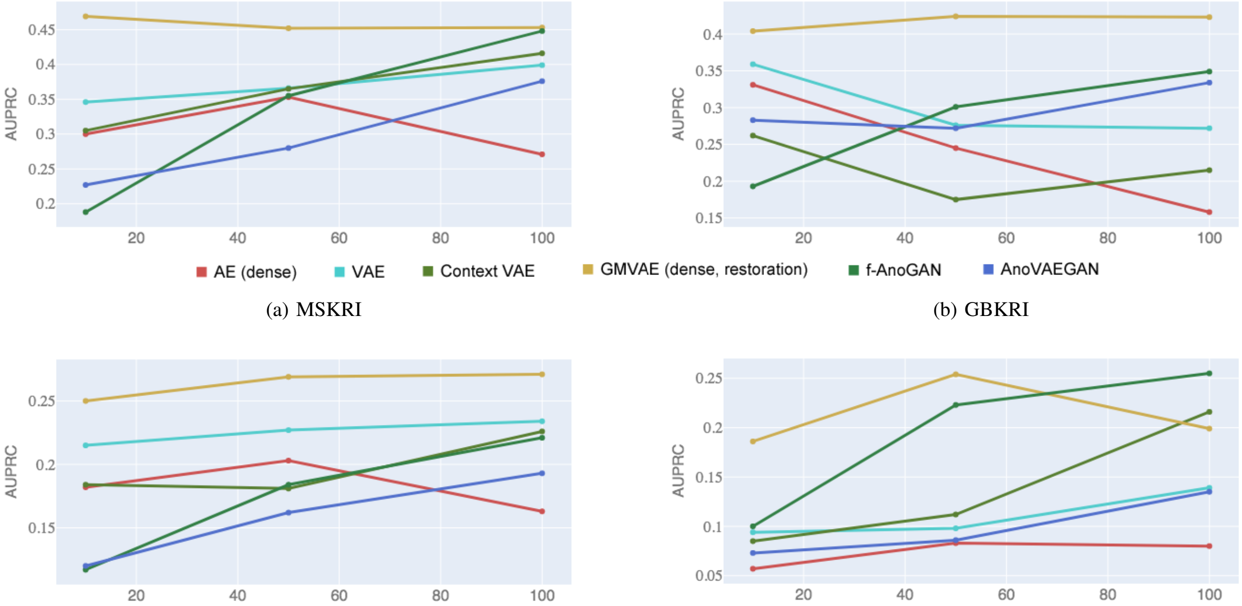
On both Multiple Sclerosis ($\mathcal{D}_{MS}$) and Glioblastoma ($\mathcal{D}_{GB}$), the restoration-based approaches with dense bottleneck constitute the top-performers, delivering results in roughly the same league. Similarly, lowest performances can be seen from the spatial models, the gradient-based UAD approach and the standard AE. However, in contrast to $\mathcal{D}_{MS}$, on $\mathcal{D}_{GB}$ there is a large performance gap between the top-performing restoration approaches and any other methods: the GAN-based methods f-AnoGAN and AnoVAEGAN drop by 10% and 4%, respectively, the performance of the VAE models degrades by at least 12% and the Constrained AE even loses 20% in AUPRC. Interestingly, the Constrained AAE gains by 10%. Multiple factors lead to the lower performance: In contrast to MS lesions, tumors do not purely appear hyperintense in FLAIR MRI. Some compartments of the tumor also resemble normal tissue, and the investigated UAD approaches have difficulties to properly delineate those. Second, tumors often are not only larger than MS lesions, but can have very complex shape (see Fig. 6d). This is hard to segment with precision—even among human annotators, there is variation.

N. How much healthy training data is enough?

In our previous experiment, we relied on 110 healthy training subjects. The question arises whether this is a sufficient amount, or if fewer scans even lead to comparable results. To give insights into the behavior of the examined models in this context, we provide a comparison of the AUPRC of conceptually most different models, all trained at varying number of healthy subjects, i.e. 10, 50 and 100% of the available training samples. Results on the four different datasets can be seen in Fig. 8. The GAN-based models, which model the healthy distribution the closest due to the Wasserstein-loss, show consistent improvements in AUPRC with a growing training set. Alone the AnoVAEGAN shows a slight drop at 50% of the training data on $\mathcal{D}_{GB}$. The overall top-performer, with one exception, is still the restoration method, here reported using the GMVAE (dense). Alone on $\mathcal{D}_{MSSEG2015}$, this GMVAE shows inconsistent behavior. Both the VAE and Context VAE, our selection from the family of VAEs with a dense bottleneck, show improved and similar performance with increasing number of training subjects on any of the MS datasets. On $\mathcal{D}_{MS}$, both models exhibit inconsistent behavior, and the VAE performs considerably better. Among all the methods, the dense AE yields the most unpredictable performance, varying greatly among different datasets and different number of healthy subjects.

O. Model Complexity

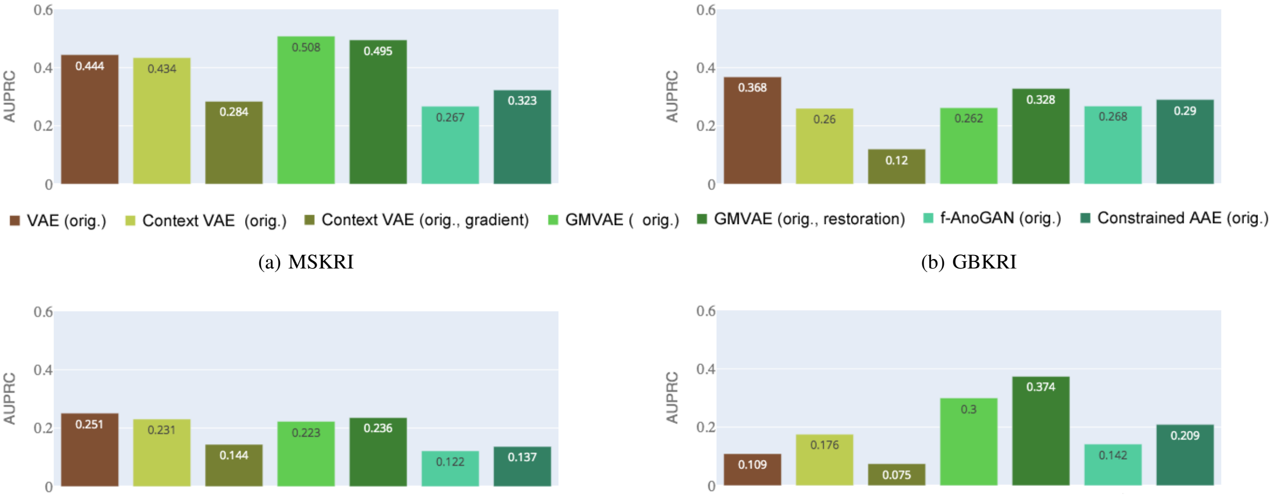
To give some insights on the relation between model complexity and segmentation performance, we further rank some of the approaches based on the architectures originally proposed in the respective papers against each other. A comparison is provided on all datasets in Fig. 9. Therein, we find the VAE and the restoration-based GMVAE methods to be stable candidates. Except for $\mathcal{D}_{MSSEG2015}$, the standard VAE approach as proposed in [7], [6], [3] shows reliable performance. Similarly, the GMVAE, especially in combination with restoration-based UAD, shows good performance across all datasets. Interestingly, the more complex VAE and Context VAE models in Fig. 9 show only comparable performance to the less complex models following our unified architecture (Fig. 5d. On $\mathcal{D}_{GB}$, none of the more complex models beat the top-performing unified restoration approach. The gradient-based approach, proposed in combination with the original Context VAE, yields lower AUPRC than its unified



(a) MSKRI (b) GBKRI

(c) MSLUB (d) MSSEG2015

Fig. 8. AUPRC of selected models trained with different numbers of healthy numbers of healthy training subjects (10, 50 and 100%, respectively).



(a) MSKRI (b) GBKRI

(c) MSLUB (d) MSSEG2015

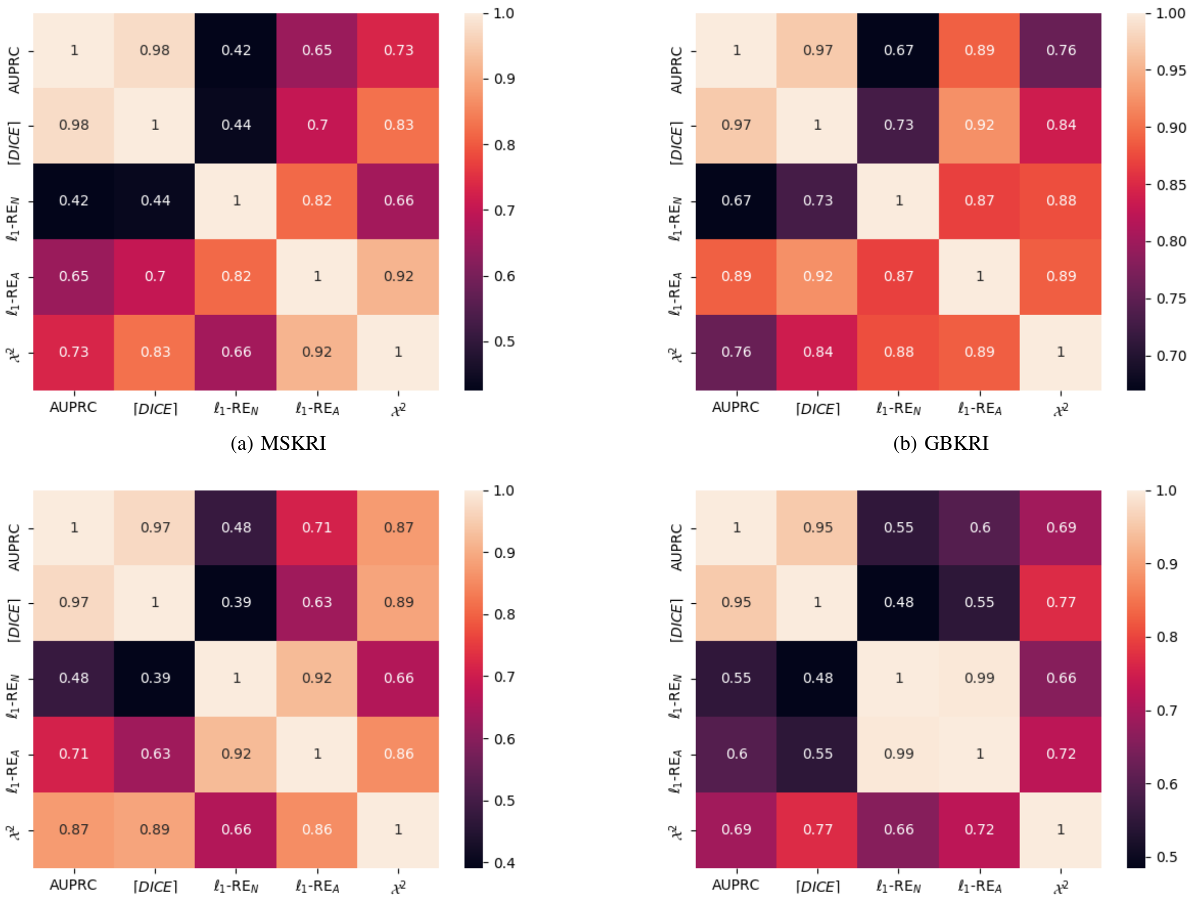
Fig. 9. AUPRC of all models and UAD approaches, using the original, more complex architectures proposed in the respective papers.

counterpart. We relate this observation to the reconstruction capabilities of models, which improve with an increase of model parameters. With increasing complexity, larger lesions such as Glioblastoma get reconstructed better as well, which is not desirable.

P. Reconstruction Fidelity and UAD Performance

From Fig. 6 it is clear that apart from spatial models, none of the approaches can reconstruct input perfectly, i.e. none of these methods leave healthy regions intact and substitute anomalous regions with plausible healthy anatomy. Nonetheless, some works perform better than others. We try to relate anomaly segmentation performance to the overlap between a models' residual histograms of normal and anomalous pixels and general reconstruction fidelity. Therefor, we correlate the AUPRC and ⌈DICE⌉ to the $\mathcal{X}^2$-distance of the aforementioned histograms, and further determine how the $\mathcal{X}^2$-distance correlates with reconstruction fidelity of normal and/or anomalous tissue. We do this for every dataset separately to find out if the correlation differs across datasets and pathologies. Fig. 10 shows the correlation heatmaps of aforementioned measures on all datasets.

On $\mathcal{D}_{MS}$ and $\mathcal{D}_{MSLUB}$, AUPRC and ⌈DICE⌉ show moderate to strong correlation to the reconstruction error on anomalous pixels ℓ_1-RE$_A$, but not so much to residuals of normal intensities ℓ_1-RE$_N$. Their correlation to the $\mathcal{X}^2$-distance



(a) MSKRI

(b) GBKRI

(c) MSLUB

(d) MSSEG2015

Fig. 10. Correlation matrices among segmentation performance, reconstruction fidelity and overlap among residual histograms of normal and anomalous intensities.

among residual histograms is the strongest. There is also a strong correlation between $\mathcal{X}^2$ and ℓ_1-RE$_A$, the correlation to ℓ_1-RE$_N$ is less pronounced. From these results we deduce that actual reconstruction fidelity is less important for UAD than clearly distinguishable residual histograms of normal and anomalous intensities.

For $\mathcal{D}_{GB}$, similar, but generally stronger correlations can be seen. Interestingly, there is also a moderate to strong, positive relationship between segmentation performance and magnitude of normal residuals. This indicates that with increasing reconstruction error on both normal and anomalous intensities, segmentation performance improves. We hyptothesize that models which reconstruct data well, also reconstruct tumors well. Models with generally poor reconstruction capabilities substitute tumors with poor reconstructions of healthy tissue, leading to better separability between anomalies and normal intensities.

On $\mathcal{D}_{MSSEG2015}$, the previously noticed correlations are hardly present. Instead, ℓ_1-RE$_N$ and ℓ_1-RE$_A$ are strongly correlated and seem to correlate similarly with all other metrics. This clearly reflects the poor contrast in the underlying MR images, which renders UAD unsuitable.

Q. Discussion

Ranking—The clear winner of this comparative study is the restoration method applied to a VAE (VAE (restoration)), which achieves best performance on $\mathcal{D}_{MS}$ and $\mathcal{D}_{GB}$, i.e. works best on different pathologies, but also achieves best performance on $\mathcal{D}_{MSLUB}$, i.e. under domain shift. However, there is a downside to the restoration method, namely runtime. A restoration of a single axial slice in 500 iterations takes multiple seconds, which for an entire MR volume accumulates quickly to multiple minutes. The feed-forward nature of purely reconstruction-based approaches allows for a much faster inference. In this context, a very promising method is the reconstruction-based f-AnoGAN, which achieves best performance on the very challenging MSSEG2015 Dataset, and is only slightly inferior to the winning restoration approach on all other datasets. Also, we find that latent variable models perform better in anomaly segmentation than classic AEs. Their reconstructions tend to be more blurry, but the gap between reconstruction errors of normal and anomalous pixels is considerably higher and allows to discriminate much better between anomalies and normal tissue. Among the latent variable models, we find the VAE to be the recommended choice, as it not only performs the best, but is the easiest to optimize. It involves fewer hyperparameters than the other approaches and does not require a discriminator network, which is a critical building block in GANs.

Open Problems—Despite all the recent successes of this paradigm, there are many questions yet to be answered. A key question is how to choose an Operating Point at which the continuous output i) can be binarized and a segmentation can be obtained or ii) an input sample can be considered anomalous. Most of the methods currently either rely on a held-out validation set to determine a threshold for binarization, or make use of heuristics on the intensity distribution. One such heuristic uses the 98th percentile of healthy data as a threshold, above which every value is considered an outlier [3]. It is necessary that more principled approaches for binarization are developed.

Although reconstruction fidelity here is far from perfect, the reviewed methods seem to be indeed capable of segmenting different kinds of anomalies. Nonetheless, we believe that the community should still aim for higher levels of fidelity and modeling MRI also at higher resolution to facilitate segmentation of particularly small brain lesions (e.g. MS lesions, which can become very small) and enhance precision of anomaly localization.

Another obvious downside of the reviewed methods is the necessity of a curated dataset of healthy data. It is debatable whether such methods can actually be called unsupervised or should be seen as weakly-supervised. The community should aim for methods which can be trained from all kinds of samples, even data potentially including anomalies, without the need for human ratings. You et al. [10] made an initial attempt towards this direction by using a percentile-based heuristic on the training data to mask out potential outliers during training, and with so called *discriminative reconstruction autoencoders* [24] an interesting concept has recently been proposed in the Computer Vision field. All in all, more research in this direction is heavily encouraged.

Generally, the field of Deep Learning based UAD for brain imaging is rapidly growing, and without the availability of a well defined benchmark dataset the field becomes increasingly confusing. This confusion primarily arises from the different datasets used in these works, which come at different resolutions, with different lesion load and different pathologies. All of these properties make it hard to compare methods. Here, we try to give an overview of recent methods, bring them into a shared context and establish comparability among them by leveraging the same data for all approaches. Nonetheless, even the datasets used in this comparative study are limited and many open questions have to remain unanswered. Since UAD methods aim to be general, they need to be evaluated on the most representative dataset possible. Ideally, a benchmark dataset for UAD in brain MRI should comprise a vast number of healthy subjects as well as different pathologies from different scanners, covering the genders and the entire age spectrum.

To date, different works do not only employ different datasets, but also report different metrics. In addition to the benchmark, a clear set of evaluation metrics needs to be defined to facilitate comparability among methods.

Last, the majority of approaches relies on 2D slices, but 3D offers greater opportunity and more context.

IV. Conclusion

In summary, we presented a thorough comparison of autoencoder-based methods for anomaly segmentation in brain MRI, which rely on modeling healthy anatomy to detect abnormal structures. We find that none of the models can perfectly reconstruct or restore healthy counterparts of potentially pathological input samples, but different approaches show different discrepancies between reconstruction-error statistics of normal and abnormal tissue, which we identify as the best indicator for good UAD performance.

To facilitate comparability, we relied on a single unified architecture and a single image resolution. The entire code behind this comparative study, including the implementations of all methods, pre-processing and evaluation pipeline will be made publicly available and we encourage authors to contribute to it. Authors might benefit from a transparent ranking which they can report in their work without having to reinvent the wheel to run extensive comparisons against other approaches.

In our discussion, we also identify different research directions for future work. Comparing different model-complexities, their correlation with reconstruction quality and its effect on anomaly segmentation performance is another research direction orthogonal to our investigations. Determining the correlation between image resolution and UAD performance is also an open task. However, our main proposal is the creation of a benchmark dataset for UAD in brain MRI, which involves many challenges by itself, but would be very beneficial to the entire community.

Acknowledgment

The authors would like to thank their clinical partners at Klinikum rechts der Isar, Munich, for generously providing their data. S.A. was supported by the PRIME programme of the German Academic Exchange Service (DAAD) with funds from the German Federal Ministry of Education and Research (BMBF).

References

[1] M. A. Bruno, E. A. Walker, and H. H. Abujudeh, “Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction,” *Radiographics*, vol. 35, no. 6, pp. 1668–1676, 2015.

[2] A. Taboada-Crispi, H. Sahli, D. Hernandez-Pacheco, and A. Falcon-Ruiz, “Anomaly detection in medical image analysis,” in *Handbook of Research on Advanced Techniques in Diagnostic Imaging and Biomedical Applications*. IGI Global, 2009, pp. 426–446.

[3] C. Baur, B. Wiestler, S. Albarqouni, and N. Navab, “Deep autoencoding models for unsupervised anomaly segmentation in brain mr images,” *arXiv preprint arXiv:1804.04488*, 2018.

[4] H. E. Atlason, A. Love, S. Sigurdsson, V. Gudnason, and L. M. Ellingsen, “Unsupervised brain lesion segmentation from mri using a convolutional autoencoder,” in *Medical Imaging 2019: Image Processing*, vol. 10949. International Society for Optics and Photonics, 2019, p. 109491H.

[5] X. Chen and E. Konukoglu, “Unsupervised detection of lesions in brain mri using constrained adversarial auto-encoders,” *arXiv preprint arXiv:1806.04972*, 2018.

[6] D. Zimmerer, F. Isensee, J. Petersen, S. Kohl, and K. Maier-Hein, "Unsupervised anomaly localization using variational auto-encoders," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 289–297.

[7] D. Zimmerer, S. A. Kohl, J. Petersen, F. Isensee, and K. H. Maier-Hein, "Context-encoding variational autoencoder for unsupervised anomaly detection," *arXiv preprint arXiv:1812.05941*, 2018.

[8] N. Pawlowski, M. C. Lee, M. Rajchl, S. McDonagh, E. Ferrante, K. Kamnitsas, S. Cooke, S. Stevenson, A. Khetani, T. Newman *et al.*, "Unsupervised lesion detection in brain ct using bayesian convolutional autoencoders," 2018.

[9] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery," in *International Conference on Information Processing in Medical Imaging*. Springer, 2017, pp. 146–157.

[10] S. You, K. C. Tezcan, X. Chen, and E. Konukoglu, "Unsupervised lesion detection via image restoration with a normative prior," in *Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning*, ser. Proceedings of Machine Learning Research, M. J. Cardoso, A. Feragen, B. Glocker, E. Konukoglu, I. Oguz, G. Unal, and T. Vercauteren, Eds., vol. 102. London, United Kingdom: PMLR, 08–10 Jul 2019, pp. 540–556. [Online]. Available: http://proceedings.mlr.press/v102/you19a.html

[11] D. Sato, S. Hanaoka, Y. Nomura, T. Takenaga, S. Miki, T. Yoshikawa, N. Hayashi, and O. Abe, "A primitive study on unsupervised anomaly detection with an autoencoder in emergency head ct volumes," in *Medical Imaging 2018: Computer-Aided Diagnosis*, vol. 10575. International Society for Optics and Photonics, 2018, p. 105751P.

[12] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *International Conference on Learning Representations*, 2014. [Online]. Available: https://arxiv.org/abs/1312.6114

[13] A. Makhzani, J. Shlens, N. Jaitly, and I. Goodfellow, "Adversarial autoencoders," in *International Conference on Learning Representations*, 2016. [Online]. Available: http://arxiv.org/abs/1511.05644

[14] N. Dilokthanakul, P. A. Mediano, M. Garnelo, M. C. Lee, H. Salimbeni, K. Arulkumaran, and M. Shanahan, "Deep unsupervised clustering with gaussian mixture variational autoencoders," *arXiv preprint arXiv:1611.02648*, 2016.

[15] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 2672–2680. [Online]. Available: http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf

[16] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," *arXiv preprint arXiv:1512.09300*, 2015.

[17] T. Schlegl, P. Seeböck, S. M. Waldstein, G. Langs, and U. Schmidt-Erfurth, "f-anogan: Fast unsupervised anomaly detection with generative adversarial networks," *Medical image analysis*, vol. 54, pp. 30–44, 2019.

[18] Ž. Lesjak, A. Galimzianova, A. Koren, M. Lukin, F. Pernuš, B. Likar, and Ž. Špiclin, "A novel public mr image dataset of multiple sclerosis patients with lesion segmentations based on multi-rater consensus," *Neuroinformatics*, vol. 16, no. 1, pp. 51–63, 2018.

[19] A. Carass, S. Roy, A. Jog, J. L. Cuzzocreo, E. Magrath, A. Gherman, J. Button, J. Nguyen, F. Prados, C. H. Sudre *et al.*, "Longitudinal multiple sclerosis lesion segmentation: resource and challenge," *NeuroImage*, vol. 148, pp. 77–102, 2017.

[20] T. Rohlfing, N. M. Zahr, E. V. Sullivan, and A. Pfefferbaum, "The SRI24 multichannel atlas of normal adult human brain structure," *Human Brain Mapping*, vol. 31, no. 5, pp. 798–819, Dec. 2009.

[21] J. E. Iglesias, C.-Y. Liu, P. M. Thompson, and Z. Tu, "Robust Brain Extraction Across Datasets and Comparison With Publicly Available Methods," *IEEE Transactions on Medical Imaging*, vol. 30, no. 9, pp. 1617–1634, 2011.

[22] J. A. Sethian, *Level set methods and fast marching methods: evolving interfaces in computational geometry, fluid mechanics, computer vision, and materials science*. Cambridge university press, 1999, vol. 3.

[23] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. International Convention Centre, Sydney, Australia: PMLR, 06–11 Aug 2017, pp. 214–223. [Online]. Available: http://proceedings.mlr.press/v70/arjovsky17a.html

[24] Y. Xia, X. Cao, F. Wen, G. Hua, and J. Sun, "Learning discriminative reconstructions for unsupervised outlier removal," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1511–1519.

APPENDIX A

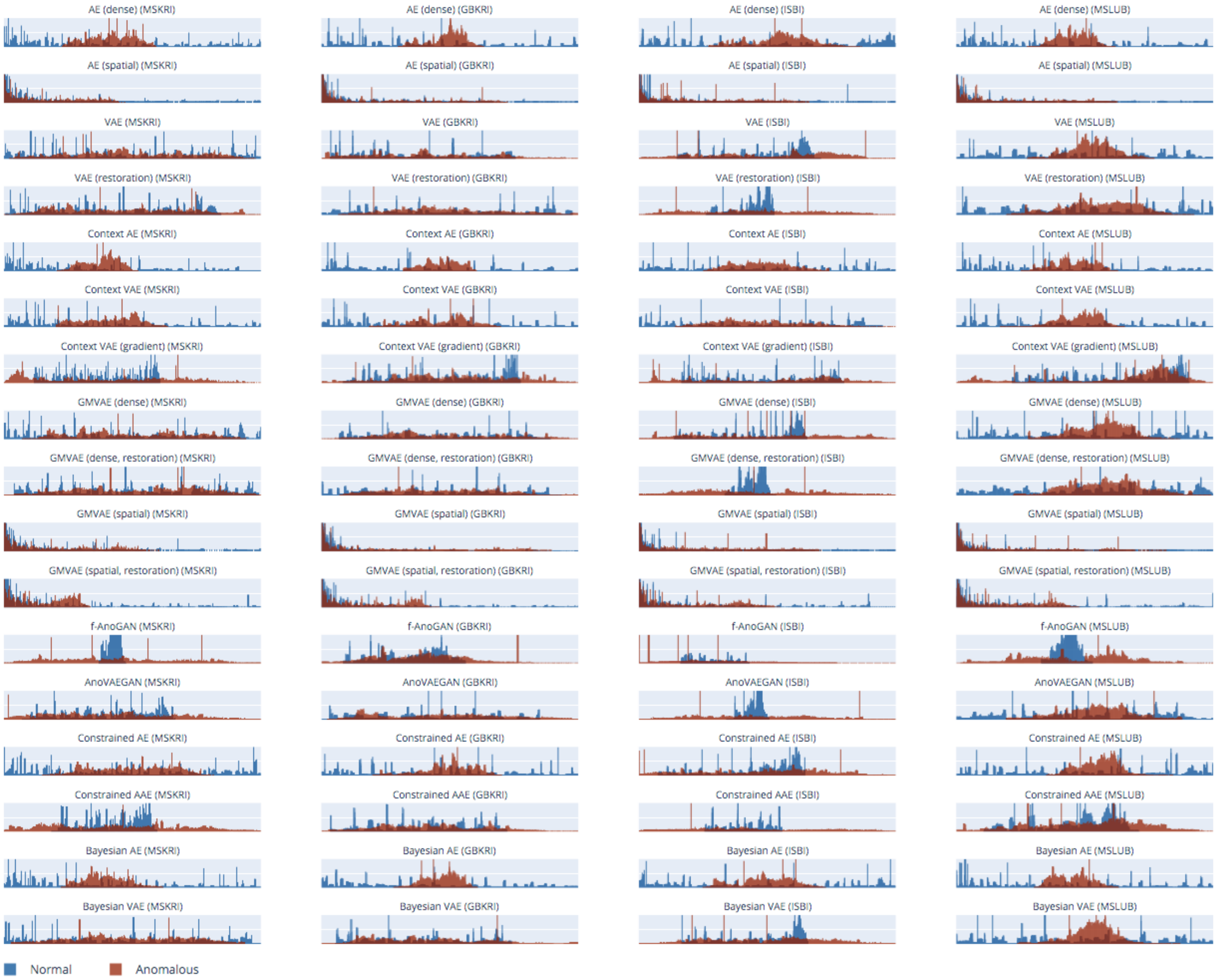


Fig. 11. Normalized histograms of residuals of normal (blue) and anomalous (red) pixels in the intensity range $\in]0; 1.0]$ (ignoring residuals which are completely 0)

3.5 Towards Self-Taught Anomaly Detection

3.5.1 Concept

Unsupervised approaches are not explicitly optimized to solve AD tasks, but rely on surrogate generative modeling and reconstruction objectives to achieve their intended task, i.e. the exploitation of their incapability to reliably reconstruct anomalous data. It may be hypothesized that the lack of an explicit optimization renders them inferior to supervised models. Supervised DL, in turn, requires large quantities of labeled data to function well, but such data is not abundant. Hybrid concepts combine the best of both worlds into cascades of unsupervised and supervised models in different ways. The concept in [16] for AD in brain MRI relies on such a cascade and a self-teaching process inspired by pseudo-labelling, showing that the combination outperforms each single entity. A similar approach has also been made in the unrelated domain of optical network intrusion detection [29]. Pseudo-labelling is a simple, yet effective method prevalent in semi-supervised DL frameworks and involves (pre-trained) supervised models to generate predictions on unlabeled samples. Together with the predicted labels and actual annotated training data, the samples are utilized to refine the supervised models. It is assumed that these pseudo-labels provide a beneficial training signal, despite their potential imperfections. In [16, 29] these pseudo-labels are provided by an unsupervised model. Closely related, Astaraki et al. [5] propose a hybrid method for lung cancer nodule detection. Therein, a supervised lung cancer nodule detection model is improved by providing it not only with the data at question, but a “healthy” appearance of that sample obtained from an autoencoding model that has been trained solely on healthy data.

3.5.2 Contribution: Fusing unsupervised and supervised deep learning for white matter lesion segmentation

Previously presented methods and contributions introduce interesting and promising solutions for UAD and UAS in brain imaging. Yet, none of the methods come close to their pathology-specific, supervised rivals. Supervised DL still outperforms the unsupervised methods by a considerable margin because of its explicit optimization for a given task. To close this gap and simultaneously alleviate the need for labeled data, this work presents a fusion of unsupervised and supervised methods and validates the method on the challenging task of WML segmentation. The approach consists of a cascade of deep representation learning for UAS and a supervised segmentation network. The unsupervised model provides labels for unlabeled data by means of its predictions which allows the data to be utilized in the subsequent training of a supervised segmentation network from the originally unlabeled samples. It is shown that this framework considerably outperforms the UAS baseline (i.e. the first component of the proposed cascade), although it is trained from the predictions made by the very same UAS model. Furthermore, the concept is also used in a semi-supervised setting, where the supervised segmentation model is trained from a small amount of data with ground-truth annotations and larger quantities of unlabeled data with labels predicted by the UAS approach. While the additional unlabeled data with its noisy “pseudo-labels” does not improve over the model trained only from labeled data, it is shown that this method allows to dramatically improve the generalization capabilities of the supervised segmentation network to data from other scanners.

Contributions|The author of this thesis was responsible for the main idea of cascading unsupervised and supervised methods for WML segmentation, for the design and implementation of the proposed framework and the evaluation pipeline, planning of the experiments, data preparation and the writing of the manuscript. Benedikt Wiestler contributed with discussions and feedback on the main ideas of the paper, as well as with datasets and their preparation. Shadi Albarqouni contributed with discussions and feedback on the main ideas of the paper, experimental design and evaluation, and was involved in proof-reading. Nassir Navab contributed with discussions and feedback on the main ideas of the paper and was involved in proof-reading.

Copyright statement|This contribution was originally published as: "**Christoph Baur**, Benedikt Wiestler, Shadi Albarqouni, and Nassir Navab. "Fusing unsupervised and supervised deep learning for white matter lesion segmentation." *In International Conference on Medical Imaging with Deep Learning, pp. 63-72. 2019*.".

Reprint Denied

The reprint of this publication was rejected on open-access platforms. The publication can be found at (http://proceedings.mlr.press/v102/baur19a.html). The details are provided below.

Publication [16]

Authors: Christoph Baur, Benedikt Wiestler, Shadi Albarqouni, Nassir Navab
Title: “Fusing Unsupervised and Supervised Deep Learning for White Matter Lesion Segmentation”
In: Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning
DOI: -

Discussion and Outlook 4

4.1 Potential clinical impact

Indisputably, neither supervised nor unsupervised DL provide all the solutions on their own towards tangible AD. Supervised methods perform well on specific tasks, but do not naturally capture the notion of anomalies. Unsupervised frameworks may be able to spot anomalies, however at reduced delineation precision and with—to some extent—even elusive behavior. However, the different paradigms may be combined and synergies may be leveraged: In analogy to anamnesis—a step-by-step process carried out by the physician to gradually narrow down the cause for the patient's condition—different DL approaches may be cascaded to determine and refine a diagnosis in Computer-Aided Diagnosis (CAD) systems. As a generalist, an anomaly detection & localization framework based on unsupervised DL may be used to determine the presence and rough location of anatomical abnormalities. The second step of the pipeline would try to identify the type of anomaly and the underlying disease. Such an algorithm doesn't have to be solely contingent on image features. The fusion of multi-modal data in the form of meta- and prior information, e.g. obtained from anamnesis or vital parameter measurements, with imaging data could be beneficial. Graph Convolutional Networks (GCNs) have shown promising results in bringing such multi-modal information together to improve classification and segmentation tasks. As a third step, a presumably supervised model would provide precise, refined delineations and facilitate quantification of specific types of anomalies for detailed diagnosis and treatment planning.

4.2 Translation to other modalities

The proposed methods for AD are not restricted to MRI and brain image analysis. Related concepts have seen preliminary investigations on head CT to detect traumatic brain injuries [85], and a hybrid AD approach for nodule detection in lung CT images has been proposed [5]. However, none of the approaches have made universal claims nor validated their performance on a sufficient variety of different modalities and parts of the anatomy.

4.3 Research challenges for UAD

In the following, key challenges for AD are identified and briefly discussed.

Taining Data | The majority of presented UAD works rely on modeling healthy anatomy of the brain. As such, a curated set of healthy MRI scans is required. Moreover, the crafting of such a dataset involves costly expert resources, albeit no pixel-level labelling is required. However, it is not obvious how many samples of normal anatomy are required to train generative models which adequately generalize to the real population. A first set of investigations in this direction has been made in [11].

Generalization|Supervised DL approaches are known to be susceptible to domain shift, which is particularly problematic in MR imaging due to the large scanner variability. Unsupervised methods such as AEs, VAEs or GANs might also be prone to this issue. Thorough investigations of the generalization capabilities of those methods, also with respect to the type of image generation (reconstruction- vs restoration-based approaches), are yet to be undertaken. Methods that do not suffer from performance degradations under domain shift are preferable as models might then be shared among different sites and no per-scanner or per-site specialist models would be required. Alternatively, the recently trending field of federated DL applied to unsupervised DL could resolve this potential issue by building a cross-site, aggregate generalist model. One may hypothesize that the presented contribution of scale-space AEs could also offer higher levels of domain invariance, as the models rely on a presumably more domain agnostic input representation in the form of residuals. This hypothesis is yet to be validated.

Benchmark|The fact that the different AD literature report their performances i) on different datasets ii) at different image resolutions on iii) different parts of the anatomy and iv) different pathologies in v) different modalities compromises comparability among the methods, makes it difficult to draw conclusions and to provide recommendations. A first attempt to enhance comparability has been made in [11]. However, a comprehensive benchmark which covers all of the aforementioned traits in different manifestations is indispensable. Such a benchmark for medical imaging is currently being established [108].

How to determine a cut-off|Generally, the decision whether a sample is anomalous or which parts of a sample are abnormal requires binarization of the inference result. The determination of such an Operating Point (OP), i.e. the cut-off value on which the binary decision is based, is non-trivial and not solved in its entirety. The majority of proposed works so far does not focus on this particular challenge and determines an OP on a held-out validation set of abnormal samples, for which ground-truth annotation is required. Few attempts to binarize the output of AD models without relying on ground-truth have been made: [15] propose to use the 98th percentile as a statistical heuristic of the anomaly residuals obtained on the healthy training set. You et al. [102] also rely completely on the normal training data and determine a threshold which satisfies a user-defined maximum number of False Positive (FP) measured on the healthy samples. The community may need to invest additional efforts towards providing principled methods for determining this binarization cut-off. Interpretable, probabilistic model output with a limited value range could be of great value in this context.

4.4 Pre-processing, Post-processing & Philosophical Considerations

Pre- and Postprocessing of data are essential components of every DL pipeline. For the success of UAD, post-processing is particularly crucial due to the optimization for a proxy-objective rather than for a specific AD goal. Post-processing techniques applied to the residual images obtained from UAD methods comprise median- and bilateral filtering, masking with anatomical priors and connected component analysis to filter FP anomaly candidates.

Indebatably, the choice of kernels for any of those filters and parameters for the connected component analysis is crucial and may be seen as prior knowledge which can only be provided by domain experts. In this light, it is questionable whether "unsupervised" is still a valid designation: how unsupervised is unsupervised, how unsupervised can it be and how unsupervised does it have to be? This certainly is a philosophical rather than a purely technical discussion. Yet, it is a necessary one.

4.5 Conclusion

This work has outlined a path from supervised to unsupervised DL for anomaly detection in medical data and highlighted a variety of contributions along this way. Focus in these contributions was put on inexpensive methods for AD in high-resolution brain MRI data with the aim to detect and delineate particularly small pathologies such as MS lesions while alleviating the need for pixel-precise ground-truth labels. The need for no or only a limited amount of labeled data increases chances for clinical applicability and might be favorable over purely supervised methods, which are known to require excessive amounts of labeled training data. The contributions range from semi-supervised methods to completely unsupervised frameworks that only rely on a curated set of healthy training data. Promising performances have been obtained on different brain pathologies and clear steps forward towards the vision of modelling high-resolution brain MRI have been made. Albeit many milestones have been reached, the paradigm of UAD still has to answer multiple research questions, among which are i) generalization capabilities of the methods, ii) the number of healthy training subjects required to model normality and the impact on AD as well as iii) principled ways to determine a decision cut-off. In order to make statements on the universal applicability and clinical impact of the proposed works, the community needs to strive for a strong benchmark.

Bibliography

[1] S. Albarqouni, C. Baur, F. Achilles, V. Belagiannis, S. Demirci, and N. Navab. “Aggnet: deep learning from crowds for mitosis detection in breast cancer histology images”. In: *IEEE transactions on medical imaging* 35.5 (2016), pp. 1313–1321 (cit. on p. xviii).

[2] J. An and S. Cho. “Variational autoencoder based anomaly detection using reconstruction probability”. In: *Special Lecture on IE* 2 (2015), pp. 1–18 (cit. on p. 30).

[3] P. Anbeek, K. L. Vincken, M. J. van Osch, R. H. Bisschops, and J. van der Grond. “Automatic segmentation of different-sized white matter lesions by voxel probability estimation”. In: *Medical image analysis* 8.3 (2004), pp. 205–215 (cit. on p. 6).

[4] M. Arjovsky, S. Chintala, and L. Bottou. “Wasserstein Generative Adversarial Networks”. In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by D. Precup and Y. W. Teh. Vol. 70. Proceedings of Machine Learning Research. International Convention Centre, Sydney, Australia: PMLR, 2017, pp. 214–223 (cit. on p. 24).

[5] M. Astaraki, I. Toma-Dasu, Ö. Smedby, and C. Wang. “Normal appearance autoencoder for lung cancer detection and segmentation”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pp. 249–256 (cit. on pp. 55, 59).

[6] H. E. Atlason, A. Love, S. Sigurdsson, V. Gudnason, and L. M. Ellingsen. “Unsupervised brain lesion segmentation from MRI using a convolutional autoencoder”. In: *Medical Imaging 2019: Image Processing*. Vol. 10949. International Society for Optics and Photonics. 2019, 109491H (cit. on p. 30).

[7] C. Baur, S. Albarqouni, S. Demirci, N. Navab, and P. Fallavollita. “CathNets: detection and single-view depth prediction of catheter electrodes”. In: *International Conference on Medical Imaging and Augmented Reality*. Springer. 2016, pp. 38–49 (cit. on p. xviii).

[8] C. Baur, S. Albarqouni, and N. Navab. “Generating highly realistic images of skin lesions with GANs”. In: *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*. Springer, 2018, pp. 260–267 (cit. on pp. xviii, 24).

[9] C. Baur, S. Albarqouni, and N. Navab. “MelanoGANs: high resolution skin lesion synthesis with GANs”. In: *arXiv preprint arXiv:1804.04338* (2018) (cit. on p. xviii).

[10] C. Baur, S. Albarqouni, and N. Navab. “Semi-supervised deep learning for fully convolutional networks”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2017, pp. 311–319 (cit. on pp. xviii, 15, 17).

[11] C. Baur, S. Denner, B. Wiestler, S. Albarqouni, and N. Navab. “Autoencoders for Unsupervised Anomaly Segmentation in Brain MR Images: A Comparative Study”. In: *arXiv preprint arXiv:2004.03271* (2020) (cit. on pp. xvii, 30, 31, 38, 60).

[12] C. Baur, R. Graf, B. Wiestler, S. Albarqouni, and N. Navab. “SteGANomaly: Inhibiting CycleGAN Steganography for Unsupervised Anomaly Detection in Brain MRI”. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II 23*. Springer International Publishing. 2020, pp. 718–727 (cit. on pp. xvii, 29).

[13] C. Baur, F. Milletari, V. Belagiannis, N. Navab, and P. Fallavollita. “Automatic 3D reconstruction of electrophysiology catheters from two-view monoplane C-arm image sequences”. In: *International journal of computer assisted radiology and surgery* 11.7 (2016), pp. 1319–1328 (cit. on p. xviii).

[14] C. Baur, B. Wiestler, S. Albarqouni, and N. Navab. “Bayesian Skip-Autoencoders for Unsupervised Hyperintense Anomaly Detection in High Resolution Brain Mri”. In: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2020, pp. 1905–1909 (cit. on pp. xvii, 35).

[15] C. Baur, B. Wiestler, S. Albarqouni, and N. Navab. “Deep autoencoding models for unsupervised anomaly segmentation in brain MR images”. In: *International MICCAI Brainlesion Workshop*. Springer. 2018, pp. 161–169 (cit. on pp. xviii, 27, 30, 31, 33, 60).

[16] C. Baur, B. Wiestler, S. Albarqouni, and N. Navab. “Fusing Unsupervised and Supervised Deep Learning for White Matter Lesion Segmentation”. In: *Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning*. Ed. by M. J. Cardoso, A. Feragen, B. Glocker, et al. Vol. 102. Proceedings of Machine Learning Research. MLR Press. London, United Kingdom: PMLR, 2019, pp. 63–72 (cit. on pp. xvii, 55, 57).

[17] C. Baur, B. Wiestler, S. Albarqouni, and N. Navab. “Scale-Space Autoencoders for Unsupervised Anomaly Segmentation in Brain MRI”. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings*. Springer International Publishing. 2020, pp. 552–561 (cit. on pp. xvii, 37).

[18] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle. “Greedy layer-wise training of deep networks”. In: *Advances in neural information processing systems*. 2007, pp. 153–160 (cit. on p. 20).

[19] L. Berlin. “Defending the “missed” radiographic diagnosis”. In: *American Journal of Roentgenology* 176.2 (2001), pp. 317–322 (cit. on p. 1).

[20] A. Birenbaum and H. Greenspan. “Longitudinal multiple sclerosis lesion segmentation using multi-view convolutional neural networks”. In: *Deep Learning and Data Labeling for Medical Applications*. Springer, 2016, pp. 58–67 (cit. on p. 14).

[21] M. A. Bruno, E. A. Walker, and H. H. Abujudeh. “Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction”. In: *Radiographics* 35.6 (2015), pp. 1668–1676 (cit. on p. 1).

[22] M. Bui, C. Baur, N. Navab, S. Ilic, and S. Albarqouni. “Adversarial Networks for Camera Pose Regression and Refinement”. In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2019, pp. 0–0 (cit. on p. xviii).

[23] M. Bui, F. Bourier, C. Baur, F. Milletari, N. Navab, and S. Demirci. “Robust navigation support in lowest dose image setting”. In: *International journal of computer assisted radiology and surgery* 14.2 (2019), pp. 291–300 (cit. on p. xvii).

[24] A. Carass, S. Roy, A. Jog, et al. “Longitudinal multiple sclerosis lesion segmentation - Resource and challenge.” In: *NeuroImage* 148 (2017), pp. 77–102 (cit. on pp. 1, 6).

[25] A. Casamitjana, S. Puch, A. Aduriz, and V. Vilaplana. “3D Convolutional Neural Networks for Brain Tumor Segmentation: a comparison of multi-resolution architectures”. In: *International Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Springer. 2016, pp. 150–161 (cit. on p. 14).

[26] T. Che, Y. Li, A. P. Jacob, Y. Bengio, and W. Li. “Mode regularized generative adversarial networks”. In: *arXiv preprint arXiv:1612.02136* (2016) (cit. on p. 24).

[27] M. Chen, T. Kanade, D. Pomerleau, and H. A. Rowley. “Anomaly detection through registration”. In: *Pattern Recognition* 32.1 (1999), pp. 113–128 (cit. on pp. 5, 6).

[28] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. “Infogan: Interpretable representation learning by information maximizing generative adversarial nets”. In: *Advances in neural information processing systems*. 2016, pp. 2172–2180 (cit. on p. 24).

[29] X. Chen, B. Li, R. Proietti, Z. Zhu, and S. B. Yoo. “Self-taught anomaly detection with hybrid unsupervised/supervised machine learning in optical networks”. In: *Journal of Lightwave Technology* 37.7 (2019), pp. 1742–1749 (cit. on p. 55).

[30] X. Chen and E. Konukoglu. “Unsupervised Detection of Lesions in Brain MRI using constrained adversarial auto-encoders”. In: *arXiv preprint arXiv:1806.04972* (2018) (cit. on pp. 30, 31).

[31] 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 (cit. on p. 13).

[32] K. Coyne. “MRI: A guided tour”. In: *Internet: https://nationalmaglab.org/education/magnet-academy/learn-the-basics/stories/mri-a-guided-tour,[June, 20, 2020]* (2020) (cit. on p. 3).

[33] N. Dilokthanakul, P. A. Mediano, M. Garnelo, et al. “Deep unsupervised clustering with gaussian mixture variational autoencoders”. In: *arXiv preprint arXiv:1611.02648* (2016) (cit. on p. 23).

[34] J. Donahue, P. Krähenbühl, and T. Darrell. “Adversarial feature learning”. In: *arXiv preprint arXiv:1605.09782* (2016) (cit. on p. 26).

[35] P. Dvorak and B. Menze. “Structured prediction with convolutional neural networks for multimodal brain tumor segmentation”. In: *Proceeding of the multimodal brain tumor image segmentation challenge* (2015), pp. 13–24 (cit. on p. 14).

[36] D. Eigen, C. Puhrsch, and R. Fergus. “Depth map prediction from a single image using a multi-scale deep network”. In: *Advances in neural information processing systems*. 2014, pp. 2366–2374 (cit. on p. 13).

[37] M. El Azami, A. Hammers, J. Jung, N. Costes, R. Bouet, and C. Lartizien. “Detection of lesions underlying intractable epilepsy on T1-weighted MRI as an outlier detection problem”. In: *PloS one* 11.9 (2016) (cit. on p. 6).

[38] D. T. Gering, W. E. L. Grimson, and R. Kikinis. “Recognizing deviations from normalcy for brain tumor segmentation”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2002, pp. 388–395 (cit. on pp. 5, 6).

[39] E. Gibson, F. Gao, S. E. Black, and N. J. Lobaugh. “Automatic segmentation of white matter hyperintensities in the elderly using FLAIR images at 3T”. In: *Journal of Magnetic Resonance Imaging* 31.6 (2010), pp. 1311–1322 (cit. on p. 5).

[40] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, et al. “Generative Adversarial Nets.” In: *NIPS* (2014) (cit. on p. 23).

[41] L. Griffanti, G. Zamboni, A. Khan, et al. “BIANCA (Brain Intensity AbNormality Classification Algorithm): A new tool for automated segmentation of white matter hyperintensities”. In: *Neuroimage* 141 (2016), pp. 191–205 (cit. on p. 6).

[42] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis. “Learning Temporal Regularity in Video Sequences”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016, pp. 733–742 (cit. on p. 30).

[43] M. Havaei, A. Davy, D. Warde-Farley, et al. “Brain tumor segmentation with deep neural networks”. In: *Medical image analysis* 35 (2017), pp. 18–31 (cit. on p. 14).

[44] L. O. Iheme, D. Ünay, O. Baskaya, et al. “Concordance between computer-based neuroimaging findings and expert assessments in dementia grading.” In: *SIU* (2013), pp. 1–4 (cit. on p. 6).

[45] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. “Image-to-Image Translation with Conditional Adversarial Networks”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), pp. 5967–5976 (cit. on p. 25).

[46] S. Jain, D. M. Sima, and D. Smeets. “Automatic longitudinal Multiple Sclerosis lesion segmentation: MSmetrix”. In: *Longitudinal Multiple Sclerosis Lesion Segmentation Challenge, International Symposium on Biomedical Imaging*. 2015 (cit. on p. 6).

[47] J. L. W. V. Jensen et al. “Sur les fonctions convexes et les inégalités entre les valeurs moyennes”. In: *Acta mathematica* 30 (1906), pp. 175–193 (cit. on p. 22).

[48] T. Kalincik, M. Vaneckova, M. Tyblova, et al. “Volumetric MRI markers and predictors of disease activity in early multiple sclerosis: a longitudinal cohort study”. In: *PloS one* 7.11 (2012), e50101 (cit. on p. 1).

[49] K. Kamnitsas, C. Baumgartner, C. Ledig, 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 (cit. on p. 18).

[50] K. Kamnitsas, C. Ledig, V. F. Newcombe, et al. “Efficient Multi-Scale 3D CNN with Fully Connected CRF for Accurate Brain Lesion Segmentation”. In: *arXiv preprint arXiv:1603.05959* (2016) (cit. on p. 14).

[51] K. Kamnitsas, C. Ledig, V. F. Newcombe, et al. “Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation”. In: *Medical image analysis* 36 (2017), pp. 61–78 (cit. on p. 6).

[52] T. Karras, T. Aila, S. Laine, and J. Lehtinen. “Progressive Growing of GANs for Improved Quality, Stability, and Variation”. In: *International Conference on Learning Representations*. 2018 (cit. on p. 24).

[53] S. Kazeminia, C. Baur, A. Kuijper, et al. “GANs for medical image analysis”. In: *Artificial Intelligence in Medicine* (2020), p. 101938 (cit. on pp. xvii, 23).

[54] D. P. Kingma and M. Welling. “Auto-Encoding Variational Bayes”. In: *International Conference on Learning Representations*. 2014 (cit. on pp. 21, 22).

[55] B Kiran, D. Thomas, and R. Parakkal. “An Overview of Deep Learning Based Methods for Unsupervised and Semi-Supervised Anomaly Detection in Videos”. In: *Journal of Imaging* 4.2 (Feb. 2018), p. 36 (cit. on p. 30).

[56] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther. “Autoencoding beyond pixels using a learned similarity metric”. In: *arXiv preprint arXiv:1512.09300* (2015) (cit. on pp. 25, 31).

[57] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. “Gradient-based learning applied to document recognition”. In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324 (cit. on p. 10).

[58] C.-H. Lee, M. Schmidt, A. Murtha, A. Bistritz, J. Sander, and R. Greiner. “Segmenting brain tumors with conditional random fields and support vector machines”. In: *International Workshop on Computer Vision for Biomedical Image Applications*. Springer. 2005, pp. 469–478 (cit. on p. 5).

[59] D.-H. Lee. “Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks”. In: *Workshop on challenges in representation learning, ICML*. Vol. 3. 2. 2013 (cit. on p. 15).

[60] 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 (cit. on p. 12).

[61] M. Lyksborg, O. Puonti, M. Agn, and R. Larsen. “An ensemble of 2D convolutional neural networks for tumor segmentation”. In: *Scandinavian Conference on Image Analysis*. Springer. 2015, pp. 201–211 (cit. on p. 14).

[62] A. Makhzani, J. Shlens, N. Jaitly, and I. Goodfellow. “Adversarial Autoencoders”. In: *International Conference on Learning Representations*. 2016 (cit. on p. 23).

[63] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. P. Smolley. “Least squares generative adversarial networks”. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE. 2017, pp. 2813–2821 (cit. on p. 24).

[64] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber. “Stacked convolutional auto-encoders for hierarchical feature extraction”. In: *International conference on artificial neural networks*. Springer. 2011, pp. 52–59 (cit. on p. 21).

[65] B. H. Menze, A. Jakab, S. Bauer, et al. “The multimodal brain tumor image segmentation benchmark (BRATS)”. In: *IEEE transactions on medical imaging* 34.10 (2014), pp. 1993–2024 (cit. on pp. 1, 6).

[66] B. H. Menze, K. Van Leemput, D. Lashkari, M.-A. Weber, N. Ayache, and P. Golland. “A generative model for brain tumor segmentation in multi-modal images”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2010, pp. 151–159 (cit. on p. 6).

[67] F. Milletari, N. Navab, and S.-A. Ahmadi. “V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation”. In: *arXiv preprint arXiv:1606.04797* (2016) (cit. on p. 13).

[68] M. Mirza and S. Osindero. “Conditional Generative Adversarial Nets”. In: *CoRR* abs/1411.1784 (2014) (cit. on p. 24).

[69] S. Nowozin, B. Cseke, and R. Tomioka. “f-gan: Training generative neural samplers using variational divergence minimization”. In: *Advances in neural information processing systems*. 2016, pp. 271–279 (cit. on p. 24).

[70] N. Pawlowski, M. C. Lee, M. Rajchl, et al. “Unsupervised Lesion Detection in Brain CT using Bayesian Convolutional Autoencoders”. In: (2018) (cit. on pp. 30, 31).

[71] S. Pereira, A. Pinto, V. Alves, and C. A. Silva. “Brain tumor segmentation using convolutional neural networks in MRI images”. In: *IEEE transactions on medical imaging* 35.5 (2016), pp. 1240–1251 (cit. on p. 14).

[72] M. Prastawa, E. Bullitt, S. Ho, and G. Gerig. “A brain tumor segmentation framework based on outlier detection”. In: *Medical image analysis* 8.3 (2004), pp. 275–283 (cit. on pp. 5, 6).

[73] J. C. Prieto, M. Cavallari, M. Palotai, et al. “Large deep neural networks for MS lesion segmentation”. In: *Medical Imaging 2017: Image Processing*. Vol. 10133. International Society for Optics and Photonics. 2017, 101330F (cit. on p. 14).

[74] A. Radford, L. Metz, and S. Chintala. “Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks”. In: *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*. 2016 (cit. on p. 24).

[75] V. Rao, M. S. Sarabi, and A. Jaiswal. “Brain tumor segmentation with deep learning”. In: *MICCAI Multimodal Brain Tumor Segmentation Challenge (BraTS)* (2015), pp. 56–59 (cit. on p. 14).

[76] A. Rasmus, M. Berglund, M. Honkala, H. Valpola, and T. Raiko. “Semi-supervised learning with ladder networks”. In: *Advances in neural information processing systems*. 2015, pp. 3546–3554 (cit. on p. 15).

[77] 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 (cit. on pp. 13, 15).

[78] M. Rosca, B. Lakshminarayanan, D. Warde-Farley, and S. Mohamed. “Variational Approaches for Auto-Encoding Generative Adversarial Networks”. In: *arXiv.org* (June 2017). arXiv: `1706.04987v1 [stat.ML]` (cit. on p. 26).

[79] F. Rosenblatt. “The perceptron: a probabilistic model for information storage and organization in the brain.” In: *Psychological review* 65.6 (1958), p. 386 (cit. on p. 7).

[80] S. Roy, J. A. Butman, D. S. Reich, P. A. Calabresi, and D. L. Pham. “Multiple sclerosis lesion segmentation from brain MRI via fully convolutional neural networks”. In: *arXiv preprint arXiv:1803.09172* (2018) (cit. on p. 14).

[81] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. *Learning internal representations by error propagation*. Tech. rep. California Univ San Diego La Jolla Inst for Cognitive Science, 1985 (cit. on pp. 20, 21).

[82] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. “Learning representations by back-propagating errors”. In: *nature* 323.6088 (1986), pp. 533–536 (cit. on p. 10).

[83] M Sabokrou, M Fathy, and M Hoseini. “Video anomaly detection and localisation based on the sparsity and reconstruction error of auto-encoder”. In: *Electronics Letters* 52.13 (June 2016), pp. 1122–1124 (cit. on p. 30).

[84] M. Sakurada and T. Yairi. “Anomaly detection using autoencoders with nonlinear dimensionality reduction”. In: *Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis*. 2014, pp. 4–11 (cit. on p. 30).

[85] D. Sato, S. Hanaoka, Y. Nomura, et al. “A primitive study on unsupervised anomaly detection with an autoencoder in emergency head CT volumes”. In: *Medical Imaging 2018: Computer-Aided Diagnosis*. Vol. 10575. International Society for Optics and Photonics. 2018, 105751P (cit. on p. 59).

[86] T. Schlegl, P. Seeböck, S. M. Waldstein, G. Langs, and U. Schmidt-Erfurth. “f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks”. In: *Medical image analysis* 54 (2019), pp. 30–44 (cit. on p. 27).

[87] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs. “Unsupervised anomaly detection with generative adversarial networks to guide marker discovery”. In: *International Conference on Information Processing in Medical Imaging*. Springer. 2017, pp. 146–157 (cit. on pp. 26, 30, 31).

[88] P. Schmidt. “Bayesian inference for structured additive regression models for large-scale problems with applications to medical imaging”. PhD thesis. lmu, 2017 (cit. on p. 6).

[89] P. Schmidt, C. Gaser, M. Arsic, et al. “An automated tool for detection of FLAIR-hyperintense white-matter lesions in multiple sclerosis”. In: *Neuroimage* 59.4 (2012), pp. 3774–3783 (cit. on p. 6).

[90] P. Seeböck, S. Waldstein, S. Klimscha, et al. “Identifying and categorizing anomalies in retinal imaging data”. In: *arXiv preprint arXiv:1612.00686* (2016) (cit. on p. 30).

[91] P. Seeböck, S. M. Waldstein, S. Klimscha, et al. “Identifying and Categorizing Anomalies in Retinal Imaging Data.” In: *CoRR* cs.LG (2016) (cit. on p. 6).

[92] M. T. Shaban, C. Baur, N. Navab, and S. Albarqouni. “Staingan: Stain style transfer for digital histological images”. In: *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE. 2019, pp. 953–956 (cit. on p. xviii).

[93] N. Shiee, P.-L. Bazin, A. Ozturk, D. S. Reich, P. A. Calabresi, and D. L. Pham. “A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions.” In: *NeuroImage* 49.2 (2010), pp. 1524–1535 (cit. on p. 6).

[94] A. Taboada-Crispi, H. Sahli, D. Hernandez-Pacheco, and A. Falcon-Ruiz. “Anomaly detection in medical image analysis”. In: *Handbook of Research on Advanced Techniques in Diagnostic Imaging and Biomedical Applications*. IGI Global, 2009, pp. 426–446 (cit. on pp. 5–7).

[95] S. Vaidya, A. Chunduru, R. Muthuganapathy, and G. Krishnamurthi. *Longitudinal multiple sclerosis lesion segmentation using 3D convolutional neural networks*. 2015 (cit. on p. 14).

[96] S. Valverde, M. Cabezas, E. Roura, et al. “Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach”. In: *NeuroImage* 155 (2017), pp. 159–168 (cit. on pp. 6, 14).

[97] K. Van Leemput, F. Maes, D. Vandermeulen, A. Colchester, and P. Suetens. “Automated segmentation of multiple sclerosis lesions by model outlier detection”. In: *IEEE transactions on medical imaging* 20.8 (2001), pp. 677–688 (cit. on pp. 5, 6).

[98] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. “Extracting and composing robust features with denoising autoencoders”. In: *Proceedings of the 25th international conference on Machine learning*. 2008, pp. 1096–1103 (cit. on p. 20).

[99] N. Weiss, D. Rueckert, and A. Rao. “Multiple Sclerosis Lesion Segmentation Using Dictionary Learning and Sparse Coding.” In: *MICCAI* 8149.Chapter 92 (2013), pp. 735–742 (cit. on p. 6).

[100] J. Weston, F. Ratle, H. Mobahi, and R. Collobert. “Deep learning via semi-supervised embedding”. In: *Neural networks: Tricks of the trade*. Springer, 2012, pp. 639–655 (cit. on p. 15).

[101] Z. Yang, W. Cohen, and R. Salakhudinov. “Revisiting semi-supervised learning with graph embeddings”. In: *International conference on machine learning*. 2016, pp. 40–48 (cit. on p. 15).

[102] S. You, K. C. Tezcan, X. Chen, and E. Konukoglu. “Unsupervised Lesion Detection via Image Restoration with a Normative Prior”. In: *Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning*. Ed. by M. J. Cardoso, A. Feragen, B. Glocker, et al. Vol. 102. Proceedings of Machine Learning Research. London, United Kingdom: PMLR, 2019, pp. 540–556 (cit. on p. 60).

[103] J. Zhao, M. Mathieu, and Y. LeCun. “Energy-based generative adversarial network”. In: *arXiv preprint arXiv:1609.03126* (2016) (cit. on p. 24).

[104] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. “Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks”. In: *2017 IEEE International Conference on Computer Vision (ICCV)* (2017), pp. 2242–2251 (cit. on p. 25).

[105] D. Zikic, Y. Ioannou, M. Brown, and A. Criminisi. “Segmentation of brain tumor tissues with convolutional neural networks”. In: *Proceedings MICCAI-BRATS* (2014), pp. 36–39 (cit. on pp. 6, 14).

[106] D. Zimmerer, F. Isensee, J. Petersen, S. Kohl, and K. Maier-Hein. “Unsupervised Anomaly Localization using Variational Auto-Encoders”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pp. 289–297 (cit. on pp. 30, 31).

[107] D. Zimmerer, S. A. Kohl, J. Petersen, F. Isensee, and K. H. Maier-Hein. “Context-encoding Variational Autoencoder for Unsupervised Anomaly Detection”. In: *arXiv preprint arXiv:1812.05941* (2018) (cit. on pp. 30, 31).

[108] D. Zimmerer, J. Petersen, G. Köhler, et al. *Medical Out-of-Distribution Analysis Challenge*. Mar. 2020 (cit. on p. 60).

List of Figures