Automatic Detection of Pathological Regions in Medical Images

Inaugural dissertation

to
be awarded the degree of

Dr. sc. med.

presented at
the Faculty of Medicine
of the University of Basel

by
Julia Wolleb
from Lupfig AG, Switzerland

Basel, 2022

Approved by the Faculty of Medicine
on application of

Prof. Dr. Philippe C. Cattin, University of Basel – *primary advisor*
Prof. Dr. Cristina Granziera, University of Basel – *secondary advisor*
Prof. Dr. Björn Menze, University of Zurich – *external expert*
Dr. Robin Sandkühler, University of Basel – *further advisor*

Basel, the 31st of October 2022

Prof. Dr. Primo Schär
Dean

To my parents.

Contents

Acknowledgements

In the first place, my thanks go to Prof. Dr. Philippe C. Cattin for giving me the opportunity to work on this project, for his motivating guidance, openness to new ideas, and continuous support in all matters that came up during my time at the DBE. Regarding the medical aspects of this project, I want to thank Prof. Dr. Cristina Granziera for her advice whenever I asked for it, her willingness to share data, and her encouragement during tough problems. It was a pleasure to have the chance to work on real-world problems in the clinic for part of my thesis.
I am also most grateful to Dr. Robin Sandkühler for his patience in answering all possible questions, for our long discussions at the blackboard or on Zoom, and for his many inspiring new ideas. It was of priceless value to have such a supportive advisor and colleague at my side to complete this project.
A very warm thanks goes to Florentin Bieder "vis-à-vis," for chatting about all big and small thoughts crossing our minds, and for his companionship - together with Eva Schnider - during such a long way of common education. My Ph.D. experience was very interesting and enriching, which is to a large part thanks to the enormous contribution of past and current members of the Center for medical Image Analysis & Navigation: Alicia Durrer, Dr. Antal Huck, Balázs Faludi, Bruno Sempéré, Carlo Seppi, Dr. Christoph Jud, Corinne Eymann-Baier, Eva Schnider, Florentin Bieder, Florian Spiess, Lorenzo Iafolla, Madina Kojanazarova, Marek Zelechowski, Massimiliano Filipozzi, Nair von Mühlenen, Negin Sahraei, Norbert Zentai, Paul Friedrich, Peter von Niederhäusern, Peter Zhang, Philippe Valmaggia, Dr. Robin Sandkühler, Samaneh Manavi, Dr. Sara Freund, Dr. Simon Pezold, Simon Fluder, Tamás Faludi, Vincent Ochs, and beyond: Dr. Beat Fasel, Cédric Duverney, Dr. Daniela Vavrecka-Sidler, Esther Zoller, Dr. Gabriela Oser Duss, Hannah Heissler, Lorin Fasel, Ludovic Amruthalingam, Muhamed Barakovic, Murali Karnam, Prof. Dr. Mathieu Sarracanie and Prof. Dr. Najat Salameh with their lab, and Dr. Reinhard Wendler. This project was financially supported by the Novartis FreeNovation initiative and the Uniscientia Foundation.
The list of things I want to thank my friends and family for is very long and reaches from wishing me good luck on the first day of my Ph.D. to keeping me down on earth during the writing process of this thesis. A special thanks goes to my parents for their continuous support and to Martin for his patience and care.

Summary

Medical images are an essential tool in the daily clinical routine for the detection, diagnosis, and monitoring of diseases. Different imaging modalities such as magnetic resonance (MR) or X-ray imaging are used to visualize the manifestations of various diseases, providing physicians with valuable information. However, analyzing every single image by human experts is a tedious and laborious task. Deep learning methods have shown great potential to support this process, but many images are needed to train reliable neural networks. Besides the accuracy of the final method, the interpretability of the results is crucial for a deep learning method to be established. A fundamental problem in the medical field is the availability of sufficiently large datasets due to the variability of different imaging techniques and their configurations.

The aim of this thesis is the development of deep learning methods for the automatic identification of anomalous regions in medical images. Each method is tailored to the amount and type of available data. In the first step, we present a fully supervised segmentation method based on denoising diffusion models. This requires a large dataset with pixel-wise manual annotations of the pathological regions. Due to the implicit ensemble characteristic, our method provides uncertainty maps to allow interpretability of the model's decisions.

Manual pixel-wise annotations face the problems that they are prone to human bias, hard to obtain, and often even unavailable. Weakly supervised methods avoid these issues by only relying on image-level annotations. We present two different approaches based on generative models to generate pixel-wise anomaly maps using only image-level annotations, i.e., a generative adversarial network and a denoising diffusion model. Both perform image-to-image translation between a set of healthy and a set of diseased subjects. Pixel-wise anomaly maps can be obtained by computing the difference between the original image of the diseased subject and the synthetic image of its healthy representation. In an extension of the diffusion-based anomaly detection method, we present a flexible framework to solve various image-to-image translation tasks. With this method, we managed to change the size of tumors in MR images, and we were able to add realistic pathologies to images of healthy subjects.

Finally, we focus on a problem frequently occurring when working with MR images: If not enough data from one MR scanner are available, data from other scanners need to be considered. This multi-scanner setting introduces a bias between the datasets of different scanners, limiting the performance of deep learning models. We present a regularization strategy on the model's latent space to overcome the problems raised by this multi-site setting.

Zusammenfassung

Medizinische Bilder sind im klinischen Alltag ein unverzichtbares Instrument für die Erkennung, Diagnose und Überwachung von Krankheiten. Bildgebende Verfahren wie Magnetresonanz- (MR) oder Röntgenaufnahmen werden eingesetzt, um die Erscheinungsformen verschiedener Erkrankungen zu visualisieren und damit Ärzten wertvolle Informationen zu liefern. Die Analyse jedes einzelnen Bildes durch Experten ist jedoch eine mühsame Aufgabe. Deep-Learning-Methoden haben ein grosses Potenzial, diesen Prozess zu unterstützen. Allerdings werden sehr viele Bilder benötigt, um zuverlässige neuronale Netze zu trainieren. Des Weiteren ist auch die Interpretierbarkeit der Ergebnisse entscheidend für die klinische Anwendung. Ein grundlegendes Problem im medizinischen Bereich ist die limitierte Verfügbarkeit von grossen Datensätzen, da die verschiedenen Bildgebungsverfahren und Konfigurationen nicht einheitlich sind.
Das Ziel dieser Arbeit ist die Entwicklung von Deep-Learning-Methoden zur automatischen Identifizierung anomaler Regionen in medizinischen Bildern. Jede Methode ist auf die Menge der verfügbaren Daten und Annotationen zugeschnitten. Zuerst präsentieren wir eine vollständig überwachte Segmentierungsmethode, die auf denoising Diffusionsmodellen basiert. Dafür wird ein grosser Datensatz mit pixelweisen manuellen Annotationen der Pathologie benötigt. Da unsere Methode implizit Ensembles generiert, können wir Unsicherheitskarten für die Interpretierbarkeit des Modells berechnen.
Manuelle pixelweise Annotationen sind schwer zu beschaffen und auch anfällig für menschliche Voreingenommenheit. Schwach überwachte Methoden umgehen diese Probleme, indem sie allein auf Informationen auf Bildebene beruhen. Wir stellen zwei verschiedene Methoden zur Erstellung von pixelweisen Anomaliekarten vor. Dafür adaptieren wir zwei verschiedene generative Modelle, nämlich ein generatives adverserielles Netzwerk und ein denoising Diffusionsmodell. Beide führen eine Bild-zu-Bild-Übersetzung zwischen einer Gruppe gesunder und einer Gruppe kranker Probanden durch. Pixelweise Anomaliekarten werden mit der Differenz zwischen dem Originalbild des kranken Probanden und dem synthetischen Bild seiner gesunden Rekonstruktion berechnet. Als Erweiterung der diffusionsbasierten Methode präsentieren wir einen flexiblen Ansatz für verschiedene Bild-zu-Bild-Übersetzungsaufgaben. Damit konnten wir die Grösse von Tumoren in MR-Bildern verändern und realistisch aussehende Pathologien zu Bildern von gesunden Probanden hinzufügen. Schlussendlich befassen wir uns mit einem häufigen Problem bei MR-Datensätzen: Wenn nicht genügend Daten von einem MR-Scanner verfügbar sind, müssen Bilder von anderen Scannern hinzugefügt werden. So entsteht ein Bias zwischen den Bildern, welcher die Modelle limitiert. Wir entwickeln eine Methode für dieses Problem und verbessern die Generalisierbarkeit der Modelle.

Chapter 1

Introduction

In recent years, the success of machine learning, especially deep learning, has revolutionized the field of medical image analysis. Novel achievements in computer vision directly influenced the development of new methods for medical tasks such as classification, semantic segmentation, or anomaly detection. However, in medical applications, problems such as the limited availability of data and labels impose challenges on the training of deep learning methods. Moreover, medical data can be diverse and heterogeneous, and a bias introduced by different acquisition methods may limit the generalization quality of machine learning models. Those challenges are tackled in a very active field of deep learning research to support and improve medical image analysis in the clinic.

1.1 Motivation

Deep learning-based anomaly detection and segmentation for medical images have the potential to support physicians in the diagnosis of diseases and to lead the attention to the relevant parts of the anatomy. The goal of this thesis is to detect visual manifestations of pathology in medical images and to outline the affected anatomical regions, as illustrated in Figure 1.1.

Figure 1.1: The overall goal of this thesis is to develop deep learning models that learn to identify the pathological regions in medical images. We aim to provide pixel-wise maps highlighting anomalous changes.

If pixel-wise ground truth segmentations of the anomalous region are provided during training on an extensive database, such a task can be solved with fully supervised segmentation approaches. While those methods have shown an impressive performance in lesion segmentation [73, 167], interpretability of the results is often not ensured but would be of great importance in clinical applications. While neural networks are often referred to as a "black box," it is crucial to understand the model's decisions and gain insight into any uncertainties. Furthermore, fully supervised segmentation approaches depend on manual pixel-wise annotations, which are time-consuming to obtain, require expert knowledge, and may even be unavailable. Moreover, if trained on manual labels, the deep learning models learn to imitate human performance and are therefore prone to human bias. Another problem occurs if already existing structures show anomalous deformations. In this case, pixel-wise ground truth segmentations are difficult to provide since, in the worst case, all anatomical structures are deformed, rendering the whole input image anomalous. Due to all these downsides of fully supervised methods, weakly supervised anomaly detection approaches are of great interest: They circumvent these issues by using only image-level labels instead of pixel-level labels during training. Thereby, they have the potential to highlight visual manifestations of a disease that were previously not in focus.
Apart from the limited availability of ground truth labels discussed above, the limited availability of data also imposes problems on the application of deep learning algorithms in medicine. Since large datasets are hard to obtain due to data privacy or differences in the acquisition protocols or hardware, data from multi-site studies need to be considered to increase the amount of training data. This can be problematic due to a bias introduced by different acquisition settings. An illustration of this problem can be found in Figure 1.2. Here, the pixel intensities for MR images of the brain acquired with two different MR scanners are shown. The images originate from the ADNI[1] and from the Young Adult Human Connectome Project (HCP) [185] dataset and are

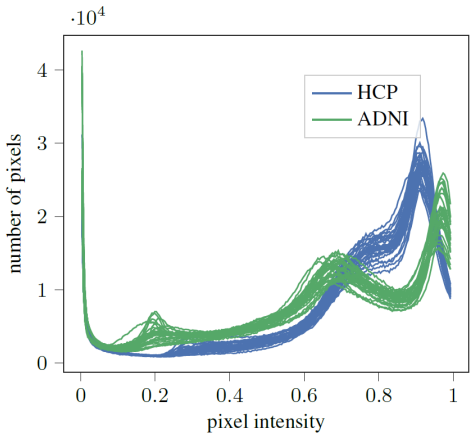


Figure 1.2: Distribution of the pixel intensities for MR images of healthy subjects of the ADNI and HCP dataset for the same MR sequence. The difference between the two datasets originates from differences in MR scanner hardware and software.

[1]Data used in preparation of this thesis were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu).

acquired using the same field strength and the same MR sequence. The difference between the two datasets, due to variations in the scanner hardware and software, can clearly be seen in the histogram. This bias disturbs the automated analysis of MR images and needs to be ignored to improve the generalization quality of deep learning methods.

1.2 Contribution

”Do I have enough data, and do I have enough labels?” – this might be the fundamental question every machine learning scientist asks first. In this project, we explore various scenarios related to the amount and type of available data that may occur in real-world applications. We present different methods with the overall goal of medical anomaly detection. In Figure 1.3, we present an overview of the different scenarios. All methods correspond to a chapter number of this thesis, each presenting a publication.

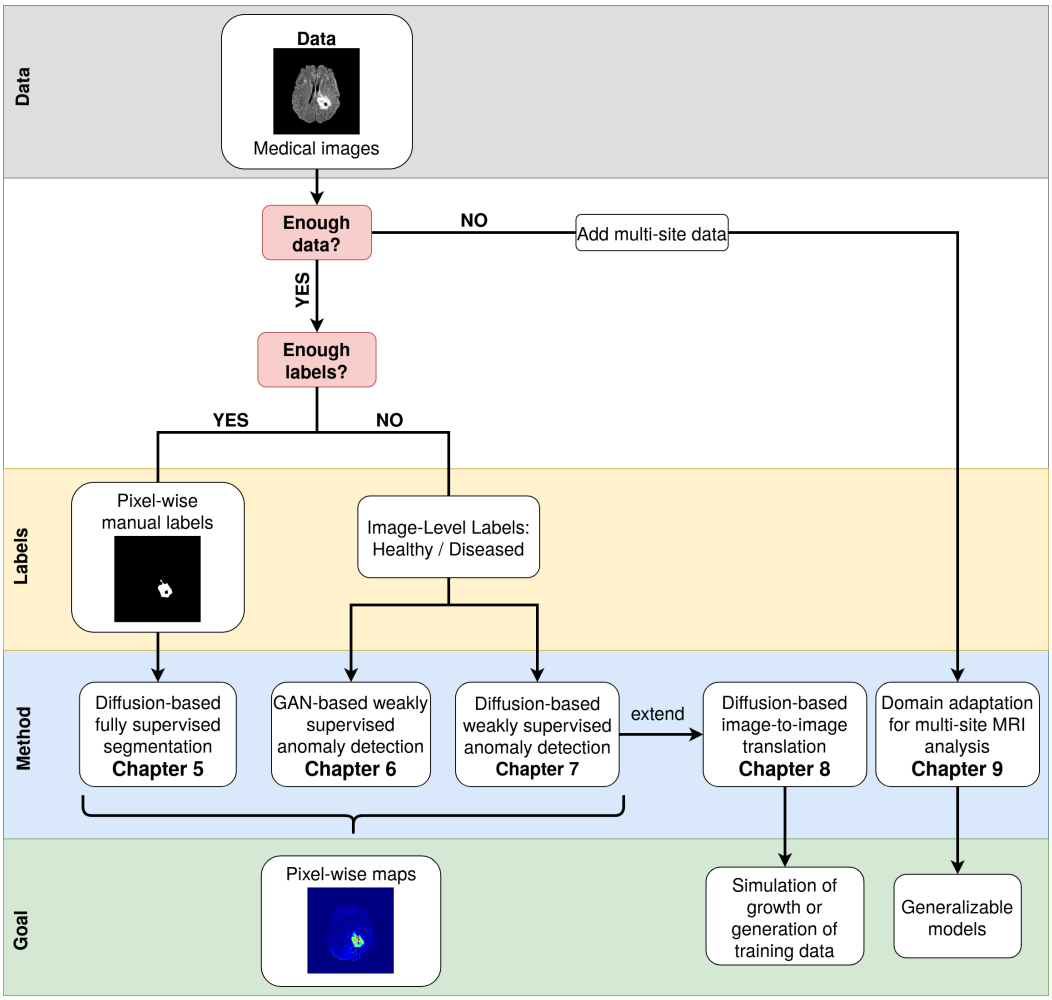


Figure 1.3: Overview of the different building blocks of this thesis.

In the first scenario, a large dataset of MR images of brains containing a tumor is available. Pixel-wise ground truth is provided for training a model to segment the brain tumor. To tackle the problem of model interpretability, we present a fully supervised segmentation method based on diffusion models that also provides pixel-wise variance maps. They can be used to measure the uncertainty of the predicted segmentation mask. This approach is presented in Chapter 5.
Next, if pixel-wise ground truth is not available, we consider anomaly detection with weak supervision. Given two unpaired sets of images, one showing healthy and one showing diseased subjects, our algorithm should automatically find the visual manifestations that make the distributions of the two datasets differ. Our method performs unpaired image-to-image translation between the two datasets. An anomaly map can then be defined by the difference between the original image of a patient and its translated synthetic healthy representation. In most previous presented methods, weakly-supervised approaches are only used to detect lesions. In contrast to that, the focus of the project presented in Chapter 6 is the detection of deformation of existing structures rather than lesions. We propose a generative adversarial network (GAN) that detects pleural effusions in lung X-ray images, which can be interpreted as a deformation of the pleural space. Furthermore, we design a synthetic dataset with pixel-wise ground truth to evaluate the performance of such anomaly detection methods.
A significant issue of GANs is their instability and cumbersome training. Therefore, another class of generative models with much more stable training, the denoising diffusion models, can be taken for image-to-image translation between healthy and diseased subjects. The reconstructed images are of very high quality and are only changed in regions showing an anomaly, resulting in very detailed anomaly maps. The straightforward training process is a significant advantage over GANs. This approach is presented in Chapter 7.
The same method can also be applied to various image-to-image translation tasks. Applying gradient guidance during diffusion-based image-to-image translation can perform a great variety of modifications. We focus on the simulation of the aging process on facial photos, brain tumor growth, and the generation of anomalous data that could be useful for evaluating other anomaly detection methods. In Chapter 8, we present this flexible framework as an extension of Chapter 7.
Only a very limited amount of training data is available for many real-world applications. We focus on a classification task on MR images and observe that using additional datasets from other MR scanners can be problematic due to the bias introduced by different scanners. The classification model tends to learn only the dominant scanner-related features rather than class-specific ones. This leads to a low generalization quality of the model. This problem is tackled in the paper presented in Chapter 9, where we introduce a method to ignore the scanner-related features by adding specific constraints on the latent space. Medical images acquired with different scanners are common in long-term or multi-center studies. Our method shows a major improvement for this scenario that can be integrated into other tasks.

1.3 Outline

Since this project comprises work on several medical image datasets, we present the different imaging modalities used to collect the datasets in Chapter 2 and provide medical background information on the diseases in focus. In Chapter 3, we provide technical details about deep-learning models. Common applications of deep learning algorithms in medicine are discussed in Chapter 4. Chapters 5 to 9 present the five publications that constitute this thesis. In Chapter 5 we present a fully supervised segmentation method based on denoising diffusion models, providing pixel-wise uncertainty maps for model interpretability. In Chapters 6 and 7, two approaches for weakly supervised anomaly detection based on GANs and diffusion models are presented. Chapter 8 builds on Chapter 7 and extends the idea to other image-to-image translation tasks. Finally, Chapter 9 presents a possible solution to the problem of limited data availability and proposes a domain adaptation method for MR harmonization across different MR scanners. We conclude by discussing the results in Chapter 10.

Chapter 2

Medical Background

As pathological changes vary significantly from disease to disease, different imaging methods are required to make them visible. Taking advantage of the physical and chemical properties of the tissues present in the body, the imaging techniques reveal information about the anatomy of the subject in focus. Section 2.1 gives a short overview of the imaging modalities used within the scope of this thesis, whereas the diseases in focus are described in Sections 2.2 to 2.4.

2.1 Medical Imaging Methods

Medical imaging is the technique used to view areas inside the human body for diagnostic or treatment purposes. Nowadays, the most common modalities used in the clinic are X-ray imaging, computed tomography (CT), magnet resonance imaging, ultrasound, positron emission tomography (PET), or single-photon emission computerized tomography [188]. In the following, the relevant imaging methods for this thesis are discussed in detail.

2.1.1 X-Ray Imaging

X-rays are a form of electromagnetic radiation with a wavelength from 10^{-8} to $10^{-12}\,\mathrm{m}$. As described in [210], they can be generated in an X-ray tube consisting of a cathode and an anode. A current is passed through the tungsten filament of the cathode and heats it up. Electrons are expelled from the filament through thermionic emission by the high energy applied. The electrons that are emitted from the cathode are then accelerated with a high voltage and hit the anode. 99% of the energy is released as heat, and 1% is emitted as two different types of X-rays: The Bremsstrahlung is generated through the deceleration of the electron and is emitted perpendicularly to the electron beam. On the other hand, when the electron collides with an inner orbit electron of an atom of the anode, both are ejected from this atom. Then, characteristic X-rays are produced when electrons change from a higher atomic orbit to a lower one in this atom. They result in peaks in the X-ray spectrum [104]. The angle of the target defines the field size of the generated X-ray beam.
For X-ray imaging, the object to be scanned is placed in-between the X-ray tube and an analog or digital image receptor such as a radiographic film or a silicon detector. This receptor

detects X-rays that pass through the object [152]. To reduce the radiation exposure, filters and collimators are placed before the object to restrict the X-ray beam and filter out low-frequency X-rays, i.e., harden the X-ray beam. To improve image quality, anti-scatter grids for filtering out scattered photons are placed after the object to prevent image blur. Sensitivity can be improved by intensifying screens [52].

Two effects dominate X-ray absorption in the object: Firstly, in the photoelectric effect, the energy is translated into the emission of electrons. Secondly, the Compton scattering is explained by the X-ray photon colliding with an electron, resulting in a scattered photon with decreased energy [152]. The absorption of the X-ray beam is proportional to the physical density of the tissue. The mass attenuation coefficient describes the absorption properties of different tissue types, determined by the material's effective atomic number and mass density [104]. Tissues with a high mass attenuation coefficient cast a shadow on the image receptor. This results in an image contrast to tissues with a lower mass attenuation coefficient. The resulting image is a two-dimensional projection of the three-dimensional object.

As different tissue types in the body differ in density and atomic structure, the mass attenuation coefficients for typical tissues also vary. As the calcium in bones absorbs X-rays the most, bones look white in the resulting image. Air absorbs X-rays the least, resulting in black areas on the image, such as the lungs. Different grayscale values in the image represent fat and soft tissue, however, soft-tissue contrast is limited [76].

In Figure 2.1, we present exemplary images of the MURA dataset [140]. As can be seen, X-ray images are a helpful tool for detecting bone fractures. Foreign objects such as screws or plates can also be visualized very well.

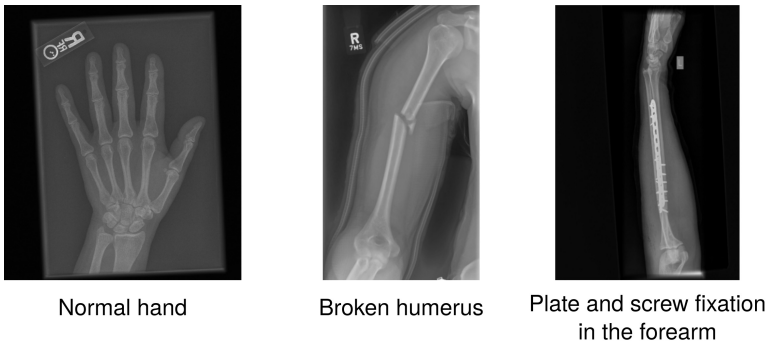


Figure 2.1: Exemplary X-Ray images of the MURA dataset, showing a healthy hand, a broken humerus, and a forearm that needed to be fixated with plates and screws.

X-ray imaging is widely used as it is cheap, non-invasive, helpful in diagnosis and medical treatment planning, and can guide the medical personnel during surgery. However, exposure to ionizing radiation increases the risk of developing cancer [86]. Moreover, certain parts of the anatomy cannot be well displayed. For example, imaging the brain is challenging, as soft tissues in general produce little contrast, but also as it is surrounded by bone that absorbs the radiation.

2.1.2 Magnetic Resonance Imaging

Magnetic Resonance Imaging (MRI) is a very commonly used imaging technology, taking advantage of the physical properties of hydrogen atoms in the human body. Hydrogen atoms themselves have a small magnetic field induced by the spin of the single proton. In the normal, relaxed state, those spins are randomly oriented. When a strong external magnetic field is applied, the protons' spins align with the direction of the external field. Using radiofrequency (RF) waves close to the so-called Larmor frequency, the magnetic moments of the protons can be excited. Those RF pulses can flip the magnetization direction. After the pulse, the spins will slowly realign to the main magnetic field, releasing energy until they are in their original state. This is denoted as the relaxation process. The longitudinal relaxation is called T1 relaxation, whereas the transverse relaxation is referred to as T2 relaxation [160]. Different tissue types have different relaxation times. During the relaxation process, energy is released in form of RF waves, which can be recorded [87].

An MR scanner is a large tube consisting of four major components: the main magnet formed by superconducting coils, gradient coils, RF coils, and a computer system [164]. In most devices for clinical applications, the main magnet has a field strength of 1.5 T or 3.0 T. This superconducting magnet consists of a series of coils wound on a cylindrical form within a bath of liquid helium for cooling. The gradient coils are placed inside the bore of an MR scanner [164]. They produce an additional magnetic field that varies in its strength along its direction and is superimposed on the main magnetic field. This allows for spatial encoding of the MR signal. Three sets of gradient coils are usually used to encode all three spatial dimensions. While the main magnetic field is kept constant once it is ramped up, the gradients need to be switched on and off quickly. The innermost ring of an MR scanner consists of the RF coils, which are used to send RF pulses and receive the signal back from the patient's body. The magnetic field produced by RF coils is perpendicular to the main magnetic field.

The MR sequence is defined by the combination of RF pulses and the gradient field, which is controlled by the computer system. The collected signals are used to calculate pixel-wise intensities using Fourier transformation, resulting in a three-dimensional output image [164]. An exemplary T1-weighted MR image of the ADNI dataset is presented in Figure 2.2, where slices of the 3D volume in the three spatial dimensions are shown. All MR scanner components mentioned above contribute to the final image. Unlike in CT imaging, only relative signal intensities are measured. Consequently, differences in the magnetic field strength [114], hardware compo-

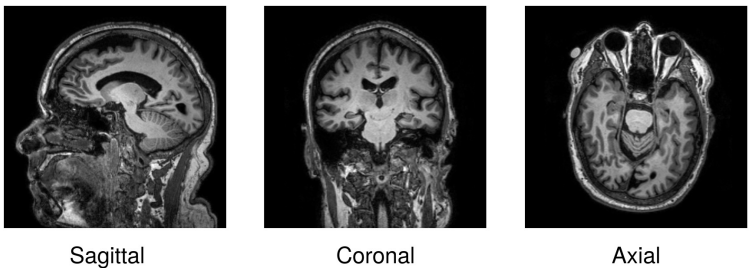


Figure 2.2: Sagittal, coronal, and axial view of an MR scan of the head. Due to the good soft-tissue contrast, the brain can be visualized in detail.

nents such as the coils [132], or software versions [40] lead to differences in the output images, making the analysis of the results prone to bias.
Two essential parameters in the acquisition of MR images are the repetition time TR, i.e., the time between successive RF pulses, and the echo time TE, which denotes the time between the delivery of the pulse and the measurement of the signal. The inversion time TI denotes the time between a 180° inversion pulse and the 90° excitation pulse in an inversion recovery pulse sequence [27]. Depending on the chosen TR and TE, we can produce different MR sequences. By choosing TI accordingly, the signal of a specific tissue type can be suppressed. The resulting images show different contrasts and can thereby be used to show different characteristics of the subject in focus. Figure 2.3 shows the three different MR sequences described below for the same subject of the BRATS2020 challenge [12, 13, 117]. The images are co-registered, interpolated to a resolution of $1\,\mathrm{mm}^3$, and the skull is removed.

T1-weighted A T1-weighted image is obtained with a short TR and a short TE [27]. The spins of the hydrogen nuclei in fat quickly realign and therefore fat appears bright on a T1-weighted image. The longitudinal magnetization realignment for water is much slower. Thus, water has a low signal and appears dark. Contrast enhancement can be achieved after bolus injection of gadolinium-based contrast agents, which shorten the T1 relaxation time [60]. In [121], a technique for three-dimensional magnetization-prepared rapid gradient-echo imaging (3D MP-RAGE) is proposed. With this sequence, the image quality and contrast between gray and white matter are improved compared with the T1-weighted spin-echo sequence [19] and thus the classification of cortical lesions in MS is improved [125].

T2-weighted A T2-weighted image is acquired using a long repetition time between RF pulses and a long signal recovery time. Water has a longer T2 relaxation time than fat. Therefore, compartments filled with water such as the cerebrospinal fluid (CSF) appear bright [23].

FLAIR Using the Fluid Attenuated Inversion Recovery (FLAIR) sequence, the signal from CSF is suppressed by using a long inversion time (TI). As for the T2-weighted images, TE and TR are chosen very long. Consequently, water appears dark compared to the T2 contrast, while abnormalities remain bright [9]. This is helpful for the evaluation of gliomas [183] and MS lesions [11].

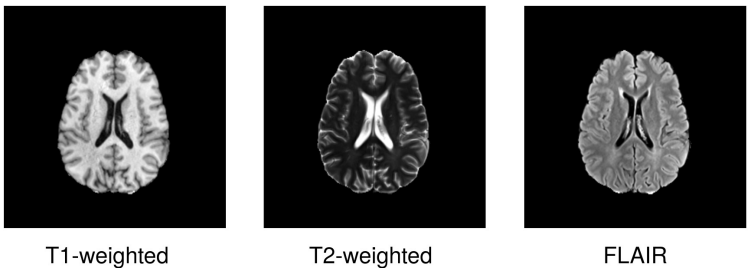


Figure 2.3: Illustration of the T1-weighted, T2-weighted and FLAIR sequences for the same slice of the same brain.

2.2 Pleural Effusions

The lungs are covered by a thin two-layered membrane called pleura. The thin space between the outer and the inner pleura is called the pleural cavity, which is filled with pleural fluid [26]. The vacuumous pleural cavity enables the expansion of the lungs, supports breathing by transmitting chest movements to the lungs, and the pleural fluid allows the layers to glide along each other during respiration [38]. In a healthy state, the production and resorption of pleural fluid are at equilibrium. Pleural effusions are defined as excess fluids in the pleural cavity due to increased production and/or decreased resorption. Transudative pleural effusions, caused by an imbalance in hydrostatic and oncotic pressure, are mainly caused by heart failure, cirrhosis, or nephrotic diseases [75]. On the other hand, exudative pleural effusions are fluid accumulations due to damage of the pleural surfaces or the capillaries. The main causes are pneumonia, cancer, gastrointestinal diseases, or tuberculosis [101]. Pulmonary embolism can be the cause of both types of pleural effusions. Symptoms include dyspnea, chest pain, and coughing [102]. Besides the effects of the underlying disease, pleural effusions may lead to lung damage, pleural thickening due to scarring, and empyema [124]. For diagnosis of pleural effusions, chest radiography is the most common technique [84], besides CT scans or ultrasound imaging. While effusions of a minimal volume of $200\,\mathrm{mL}$ can be seen with the postero-anterior view using X-ray imaging, a lateral view reveals effusions of volume $50\,\mathrm{mL}$ or larger [75].
For the treatment of pleural effusion, the underlying cause must be cured. The effusion can be drained through therapeutic thoracentesis or tube thoracostomy if respiratory problems occur. The creation of pleural sclerosis performed with sclerosing agents may prevent the recurrence of pleural effusions [84, 137]. In Figure 2.4, we give exemplary X-ray images of the CheXpert dataset [72] of a healthy subject and two diseased subjects with a right-sided and a bilateral pleural effusion, respectively.

Healthy | Right-sided pleural effusion | Bilateral pleural effusion

Figure 2.4: Lung X-ray images showing a healthy subject, a subject with a right-sided pleural effusion, and a subject with a bilateral pleural effusion.

2.3 Multiple Sclerosis

Multiple sclerosis (MS) is an autoimmune disease of the central nervous system [35], and with 2.8 million people suffering from MS worldwide, it is one of the most common neurological disorders [92]. Due to inflammatory attacks, the myelin sheath and the underlying axons are

damaged, leading to brain and spinal cord lesions [90]. Once the inflammation subsides, a scarring of the myelin insulation is likely. This scar tissue in multiple areas leads to the name *multiple sclerosis* [90]. The damage of the myelin insulation leads to disruption of the passing of the electrical signals along the axons, resulting in a disorder of sensory and motor functions. Primary symptoms include blurry vision, numbness, tingling, bladder dysfunction, and fatigue. With the progression of the disease, the effects usually worsen, leading to muscle stiffness, spasms, breathing problems, speech difficulties, bowel or bladder incontinence, and a higher risk of depression [36]. While life expectancy is about 6 to 7 years lower compared to the healthy population [153], quality of life is strongly affected [178].
One can distinguish between four types of MS [109]. The clinically isolated syndrome (CIS) is a single episode of symptoms due to inflammation and demyelination that lasts longer than 24 hours [119]. If the symptomatic phases recur, this is the relapsing-remitting MS (RRMS), where a recovery phase follows inflammatory demyelination. RRMS is the most common disease course [178]. After 8-12 years, most RRMS patients transition to secondary progressive MS (SPMS) with continuous neurological decline. A less common type is the primary progressive MS, where the disease worsens progressively without any remission phases [178].
Up to date, the causes of MS incidence are still unknown. However, a combination of hereditary and environmental factors increases the prevalence [96]. Viral infections, microbial infections, stress, lack of vitamin D, obesity, or smoking are known to increase the incidence of MS [129]. Two out of three patients are female, and young adults are disproportionally affected by MS, as the global average age of MS diagnosis is only 32 [92]. There is also a strong geographical gradient, showing that MS is more common in countries further away from the equator [92]. While MS is not inherited, there are genetic risk factors leading to an increased incidence in relatives of MS patients [129].
The diagnosis of MS can be challenging, and other diseases need to be ruled out. Tests include neurological examinations, MR scans showing local demyelination, lumbar puncture, evoked potential tests, or blood tests [21]. While there is no cure for MS, treatment can include disease-modifying therapies to reduce the number of relapses, steroids or plasma exchange during MS attacks, or symptom control through physiotherapy or medications [83]. The criteria for an MS diagnosis usually follow the *McDonalds criteria*, extensively using MRI evidence to check the occurrence of lesions at different times in different parts of the central nervous system [177].

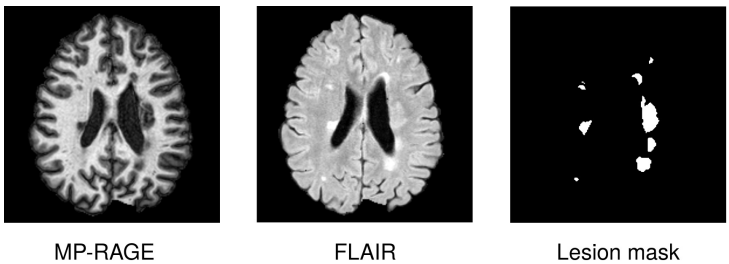
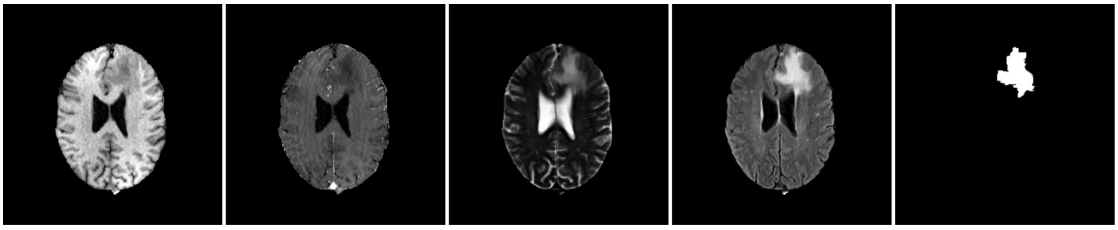


Figure 2.5: FLAIR and MP-RAGE images of an MS patient from the LMSLS challenge. On the right, the manually segmented lesion mask is presented, showing that the lesions are scattered over the brain.

In Figure 2.5, brain MR images of an exemplary MS patient of the Longitudinal Multiple-Sclerosis Lesion Segmentation (LMSLS) challenge [24] are presented. The MP-RAGE and FLAIR sequences are shown, as well as the manually segmented lesion mask.

2.4 Glioma

A glioma is the most common type of primary tumor occurring in the brain and the spinal cord and originates from glial cells. Research has shown that complex genetic, chromosomal, and epigenetic changes cause this outbreak of cancer [120]. Glioma types include astrocytomas, ependymoma, and oligodendrogliomas. Grade IV tumors, according to the World Health Organization's (WHO) classification system, are called glioblastomas [192]. They account for 60–70% of all gliomas [78], with a 5-year survival rate of only 6% [180].
In 2016, the WHO proposed a new classification scheme for tumors of the central nervous system, taking molecular alterations and histology into account [106]. This approach improved the homogeneity of the clinical outcomes of the different subgroups of gliomas and reduced misclassification.
Tumor incidence depends on risk factors such as ethnicity, age, sex, geographic location, exposure to radiation, and genetic factors. The survival rate varies with the tumor subtype, age, and sex [120]. Depending on the size and location of the tumor, common symptoms include nausea, balance difficulties, seizures, vision problems, confusion, memory loss, and personality changes [70]. For glioma diagnosis, neurological tests and imaging methods such as PET, CT, and MR imaging can be used to identify the affected brain regions and tumor size. The exact tumor type can be defined by histology with a tissue biopsy [192].
Glioma treatment options include tumor resection, radiation therapy, chemotherapy, and experimental clinical trials. However, there is a significant impact on the quality of life [191].
Figure 2.6 shows a brain MR image of the BRATS2020 challenge where the skull was removed. Four different MR sequences visualize different tumor characteristics present in the brain. A manual pixel-wise labelmask of the tumor is provided in the last image.



T1-weighted | T1-weighted with contrast enhancement | T2-weighted | FLAIR | Lesion mask

Figure 2.6: All four MR sequences and the tumor segmentation mask of an exemplary subject of the BRATS2020 dataset.

Chapter 3

Deep Learning

Artificial Intelligence (AI) is a very broadly used term nowadays, with various definitions. A century before the first computer was invented, Ada Lovelace wrote that Charles Babbage's *Analytical Engine* "might act upon other things besides numbers [...] the Engine might compose elaborate and scientific pieces of music of any degree of complexity or extent" [107]. In 1950, Alan Turing asked himself whether machines can think [179]. Later on, AI was defined as the "study of the computations that make it possible to perceive, reason and act" [193], or as "the science and engineering of making intelligent machines" [115].
Machine Learning is a subfield of AI that focuses on data-driven learning. Machine learning models automatically extract patterns from raw data during a training process. A classic example of such a model is the perceptron proposed by Rosenblatt [148, 149]. It is a basic neural network with learnable parameters that can be used to solve linear classification tasks.
Deep Learning is a subfield of machine learning that experienced a wave of success after 2010. With increasing computational power and the widespread availability of graphics processing units (GPUs), the practical employment of ideas and theories of the middle of the 20th century [116, 149] could finally be realized. Initially inspired by the function of a human brain, deep

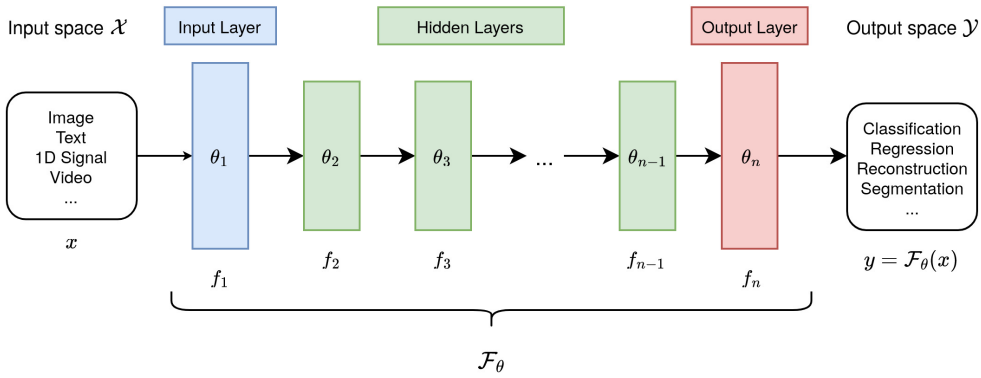


Figure 3.1: A deep learning architecture with a sequence of hidden layers. Each layer f_i represents a function defined by a set of learnable parameters θ_i.

learning methods are based on neural networks, where multiple layers are stacked to a *deep* architecture. Based on a sequence of perceptrons, the multilayer perceptron (MLP) was the first approach with a deep architecture. This allows the MLP to approximate non-linear functions. In Figure 3.1, a basic example of such a deep architecture is presented. Each layer represents a function f_i that is defined over its learnable *weights* or *parameters* θ_i. The network $\mathcal{F}_\theta$ consists of a sequence of layers $\{f_i\}_{i=1}^n$ with parameters $\theta = \{\theta_i\}_{i=1}^n$, such that the output of one layer serves as input for the next layer. The input can have various formats such as images, text, audio signals, or even videos. With the series of hidden layers, the model can extract complex features of the input data and combine them to the desired output. In general, the goal is to find an optimal parameter configuration θ^* such that the neural network $\mathcal{F}_\theta$ approximates an unknown function G. Both functions $\mathcal{F}_\theta$ and G map the input space $\mathcal{X}$ to the output space $\mathcal{Y}$:

$$\mathcal{F}_\theta : \mathcal{X} \to \mathcal{Y}, \quad G : \mathcal{X} \to \mathcal{Y} \tag{3.1}$$

For a given input x, the output is defined by

$$y = \mathcal{F}_\theta(x) = f_n(f_{n-1}(...(f_2(f_1(x)))...). \tag{3.2}$$

The goal is to adapt the learnable parameters θ such that y matches a ground truth label $\tilde{y} = G(x) \in \mathcal{Y}$ for any input $x \in \mathcal{X}$.

3.1 Training Procedure

The training of a deep neural network $\mathcal{F}_\theta$ is defined by the iterative optimization process presented in Figure 3.2. Given some training data point x, the learnable weights θ must be updated such that $y = \mathcal{F}_\theta(x)$ matches $\tilde{y} = G(x)$. Therefore, a loss function $\mathcal{L}$ is defined for comparison of y and $\tilde{y}$, dependent on the task. Following the idea of gradient descent, the parameters θ are optimized using the gradient $\nabla_\theta \mathcal{L}(y, \tilde{y})$.

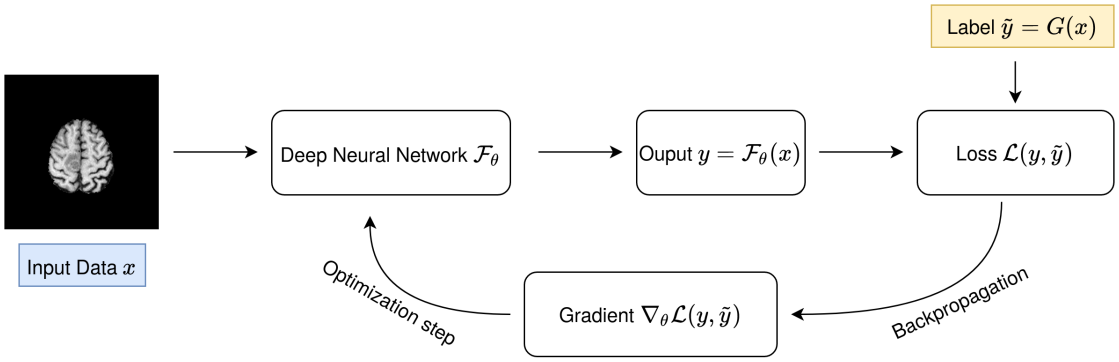


Figure 3.2: Illustration of the training process of a neural network $\mathcal{F}_\theta$. The goal is to find the optimal parameter configuration θ^* that minimizes the loss objective $\mathcal{L}$.

3.1.1 Loss functions

As shown in Figure 3.2, the loss function $\mathcal{L}$ evaluates how well the model $\mathcal{F}_\theta$ matches the target function G by commonly providing a scalar value. The model parameters are optimized to minimize this loss term. Therefore, the loss functions are specifically designed to best support the learning of the task. We present four examples that are widely used.

Mean Absolute Error The mean absolute error (MAE) is defined by the absolute distances between the output $y = \mathcal{F}_\theta(x)$ and the target label $\tilde{y} = G(x)$. P denotes the dimension of $\tilde{y}$ and y: For regression tasks such as age prediction, y and $\tilde{y}$ are scalar values, such that $P = 1$. For image reconstruction tasks, this loss is applied pixel-wise, and P is the number of pixels in y and $\tilde{y}$.

$$\mathcal{L}_{MAE}(y, \tilde{y}) = \frac{1}{P}\sum_{j=1}^{P} |y_j - \tilde{y}_j| \,. \tag{3.3}$$

Mean Squared Error The mean squared error (MSE) is defined similarly to the MAE, but the squared distance between y and $\tilde{y}$ is taken into account. This ensures that output values far from the target contribute more to the loss function.

$$\mathcal{L}_{MSE}(y, \tilde{y}) = \frac{1}{P}\sum_{j=1}^{P} \left(y_j - \tilde{y}_j\right)^2 . \tag{3.4}$$

Cross-entropy Loss For a classification task between C classes, a classification network predicts scores y_k for all classes $k \in \{1, ..., C\}$. To compute an output probability value between 0 and 1 for each class, the output of the model is passed through a softmax function, such that the sum of all probabilities $\sum_{k=1}^{C} p_{k,j} = 1$ at each entry $j \in \{1, ..., P\}$:

$$p_{c,j} = \frac{\exp(y_{c,j})}{\sum_{k=1}^{C} \exp(y_{k,j})}, \quad \text{for}\, c \in \{1, ..., C\},\ j \in \{1, ..., P\}. \tag{3.5}$$

The cross-entropy loss is then defined as

$$\mathcal{L}_{CE}(p_1, ..., p_C, \tilde{y}) = -\sum_{j=1}^{P}\sum_{k=1}^{C} \mathbb{1}_{k=\tilde{y}_j} \log p_{k,j} \,. \tag{3.6}$$

Dice Loss As an alternative to the cross-entropy loss (3.6), the Dice loss can be defined to maximize the soft Dice coefficient. This loss is a widespread objective for semantic image segmentation, where the predicted label p and the ground truth label $\tilde{y}$ have image dimension with a number of pixels P:

$$\mathcal{L}_{Dice}(p_1, ..., p_C, \tilde{y}) = 1 - \frac{1}{C}\sum_{k=1}^{C} \frac{2 * \sum_{j=1}^{P} p_{k,j}\tilde{y}_{k,j}}{\sum_{j=1}^{P} p_{k,j}^2 + \sum_{j=1}^{P} \tilde{y}_{k,j}^2}. \tag{3.7}$$

3.1.2 Parameter Optimization

Given a training dataset $\mathcal{T} = \{x_1, ..., x_n\}$, the optimization problem is given by

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} \mathcal{L}(F_\theta(x_i), G(x_i)). \tag{3.8}$$

The overall goal is to find the optimal parameter configuration θ^* such that the output of the neural network $F_\theta(x)$ matches an unknown target function $G(x)$.
This minimization problem is tackled with gradient-based learning. Gradient descent, also known as steepest descent, is a first-order iterative optimization algorithm for finding a local minimum of a differentiable function. In the best case, the function to be optimized is convex, guaranteeing convergence to the global minimum. In the iterative optimization process, the new parameter configuration θ^{t+1} is computed given the current state θ^t:

$$\theta^{t+1} = \theta^t - \gamma \nabla_\theta \mathcal{L}, \tag{3.9}$$

where the gradient $\nabla_\theta \mathcal{L}$ is the partial derivative

$$\nabla_\theta \mathcal{L} = (\frac{\partial \mathcal{L}}{\partial \theta_1}, \frac{\partial \mathcal{L}}{\partial \theta_2}, ..., \frac{\partial \mathcal{L}}{\partial \theta_q}) \quad \text{for all learnable model parameters } \theta = \{\theta_1, ..., \theta_q\}. \tag{3.10}$$

The so-called *learning rate* is denoted by γ, which is the step size in the direction of the steepest descent. As the neural network $\mathcal{F}_\theta$ consists of many interconnected layers and activation functions, the backpropagation algorithm [151], based on the chain rule, is applied to efficiently compute the gradient with respect to all parameters θ of the network.
Taking the whole training dataset into account for gradient descent, using (3.9) leads to a high computational cost. To circumvent this issue, the gradient is estimated using only a small set of samples, i.e., a mini-batch $\mathcal{B}_t = \{x_1, ..., x_m\}$ of size m randomly sampled from the training set for every optimization step t [54]. Stochastic gradient descent is defined by applying (3.9), where the loss $\mathcal{L}$ is computed only taking the mini-batch $\mathcal{B}_t$ into account.
Building on stochastic gradient descent [146], many optimization methods have been proposed, among which the Adam optimizer [93] or Root Mean Squared Propagation (RMSprop) [62] are popular.

3.1.3 Supervision Schemes

Depending on the type of ground truth information that is provided during training, one can distinguish between unsupervised, weakly-supervised, semi-supervised, fully-supervised and self-supervised learning. In unsupervised learning, no labels $\tilde{y}$ are given, and the learning can be performed based on clustering or dimension reduction methods [143]. The other extreme is full supervision, where the manual ground truth is given in full detail for all inputs. In medical applications, this could, for example, be the hand-segmented pixel-wise ground truth label of a tumor for an MR image. However, those labels require expert knowledge and are time-consuming to obtain. Moreover, there is variability between human raters, which injects a bias into the labels. A model trained with full supervision learns to imitate human performance. To circumvent these

issues, weakly supervised methods only require labels containing less information. In the medical field, this could be image-level labels indicating whether the subject in focus suffers from a specific disease or not.
Semi-supervised learning is a mixture of supervised and unsupervised methods, where the labels are only provided for a part of the training set [37, 150]. This opens the possibility of enlarging the dataset with unannotated data.
Self-supervised learning can be seen as an autonomous form of supervised learning, where the labels are auto-generated rather than human-made. A generalizable representation is learned from unlabelled data by solving a supervised proxy task, which is often unrelated to the target task [2, 64].

3.2 Components of Neural Networks

Building on the initial idea of the MLPs, modern deep neural networks are composed of different functions in each layer, optimized for memory consumption, gradient flow, and stabilized training. A standard building block consists of a convolutional or fully connected layer, followed by a normalization layer and an activation function. We present the different components in the following subsections.

3.2.1 Convolutional Layers

For image processing, the multilayer perceptron has its downsides. All model parameters are interconnected, which leads to a very high computational cost. To decrease the number of parameters, [97] proposed convolutional neural networks (CNNs). This approach is inspired by the idea that nearby pixels are correlated and describe a local feature. A combination of such local features can describe an object. The hidden layers no longer connect all parameters with each other, but a window is defined that only takes a subspace of the input into account. This window is called the *kernel*.
By sliding the kernel over the whole input, all input information is considered. An overview of the sliding window approach of the kernel is given in Figure 3.3. The kernel framed in red consists of learnable weights $w_{i,j}$. This kernel is slided over the input image I with a given stride s. A convolution defines the kernel operation. For a two-dimensional input image I with kernel K and stride s, the convolution is given by

$$S(i,j,s) = (I * K)(i,j) = \sum_m \sum_n I(is-m, js-n)K(m,n). \tag{3.11}$$

The output of the layer is given by S, of which the stride and the kernel dimension define the dimension. S is the so-called *feature map* and serves as input for the next layer of the neural network. Image dimensions are preserved if $s = 1$. Other choices of $s \in \mathbb{N}$ lead to downsampling of the image dimensions. 1D or 3D convolutions follow the same principle.
Since the parameters of the kernel are shared for all locations of the input image, similar features that appear in different parts of the input are extracted. By repeating this process over multiple layers, the model can extract complex features by locally combining the features of the previous

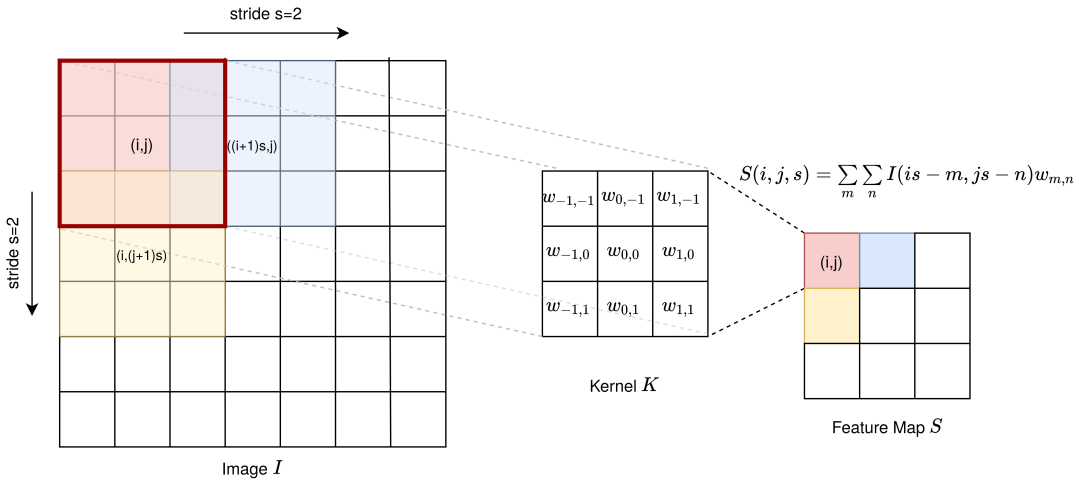


Figure 3.3: Illustration of the convolution kernel, which is slided over the input image I. In this example, the stride s is 2, and the kernel is of dimension 3×3.

layer. With this setup, three key concepts help optimize the machine learning model: sparse connectivity due to a relatively small number of weights in the kernels, equivariant representations with respect to translation, and parameter sharing [54]. By learning the weights of the kernel, a CNN learns an internal representation of an automatically selected input feature.
For each convolutional layer, there is a number of input channels a, and a number of output channels b. In Figure 3.4, the procedure for multiple input channels c_i for $i \in \{1, ..., a\}$ and multiple output channels S_j for $j \in \{1, ..., b\}$ is presented. The number of kernels in this layer is defined by b, and each kernel K_j has a channels $\{K_{j,i}\}_{i=1}^{a}$. The convolution (3.11) is computed between $K_{i,j}$ and c_i for all $i \in \{1, ..., a\}$. The results are summed up over $i \in \{1, ..., a\}$ to compute the output feature map S_j.

3.2.2 Fully Connected Layers

In fully connected layers, all entries of one layer are connected to all entries of the next layer. For an input vector $x \in \mathbb{R}^n$ and an output vector $y \in \mathbb{R}^m$, all entries of x and y are interconnected with

$$y_k = \sum_{i=1}^{n-1} w_{i,k} x_i + w_{0,k}, \quad \text{for } k \in \{1, ..., m\}, \tag{3.12}$$

where the weight matrix $W \in \mathbb{R}^{m \times n}$ consist of learnable weights $w_{i,k}$. Since interconnecting all entries of x and y with learnable weights requires a high computation power, usually fully connected layers are combined with convolutional layers in modern neural networks for image processing.

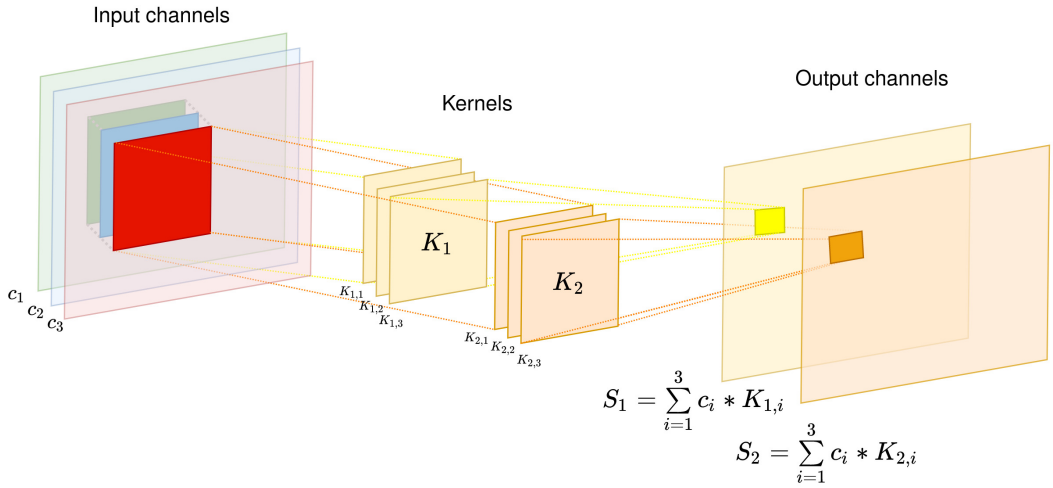


Figure 3.4: Illustration of the convolutional layer. In this example, the number of input channels a equals 3, whereas the number of kernels and therefore the number of output channels b equals 2.

3.2.3 Normalization Layers

The output of the convolutional or fully connected layer is passed to a normalization layer. A typical example is batch normalization. This normalization reduces the internal covariate shift and speeds up the training process [71]. It also has a regularizing effect, resulting in a better generalization quality of the model. When a batch $\mathcal{B} = \{x_1, ..., x_m\}$ of size m of images is passed through the model, the resulting output values are denoted as $\mathcal{F}_\theta(\mathcal{B}) = \{y_1, ...y_m\}$. Then, the batch normalization layer computes

$$BN(y_i) = \gamma \frac{y_i - \mu_\mathcal{B}}{\sigma_\mathcal{B}} + \beta \quad \forall\, i \in \{1, ..., m\}, \tag{3.13}$$

where $\mu_\mathcal{B} = \frac{1}{m} \sum_{i=1}^{m} y_i$, $\sigma_\mathcal{B}^2 = \frac{1}{m} \sum_{i=1}^{m} (y_i - \mu_\mathcal{B})^2$, and γ and β are parameters learned by the model. However, especially when dealing with 3D data and limited computation power, only small batch sizes are possible. Therefore, other normalization strategies such as instance [182], layer [10], or group normalization [200] are also regularly used. In Figure 3.5, an overview of those four common normalization layers is visualized. The color indicates over which dimension the layer is normalized to a zero mean and unit variance by applying (3.13) over the colored values. For specific applications such as style transfer, adaptive instance normalization [69] can be implemented to support the training.

3.2.4 Activation Functions

Since the functionality of the human brain inspired the first deep neural networks, the initial idea was to simulate the action potential firing in a neuron [141]. Therefore, the Heaviside

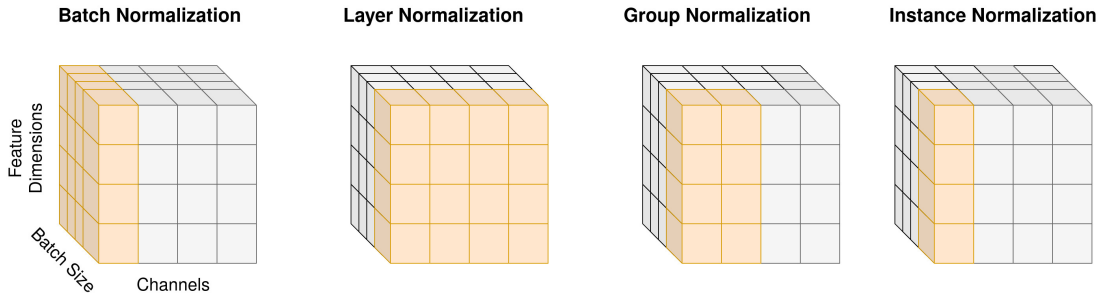


Figure 3.5: Illustration of four common normalization layers. The orange color indicates the dimensions over which the normalization function is applied.

function, a binary step function, was already used by the initial formulation of the perceptron [148]. If no activation function is applied, each layer will represent a linear function. This would hinder the mapping of more complex relations. Therefore, non-linear functions are inserted after the normalization layers, the so-called activation functions. To circumvent the problems of the zero gradient and non-differentiability in zero, other activation functions were defined [165]. In Figure 3.6, besides the Heaviside function, we plot some frequently used non-linear activation functions: the sigmoid, the hyperbolic tangent (Tanh), the rectified linear unit (ReLU), the leaky ReLU, and the softplus functions.

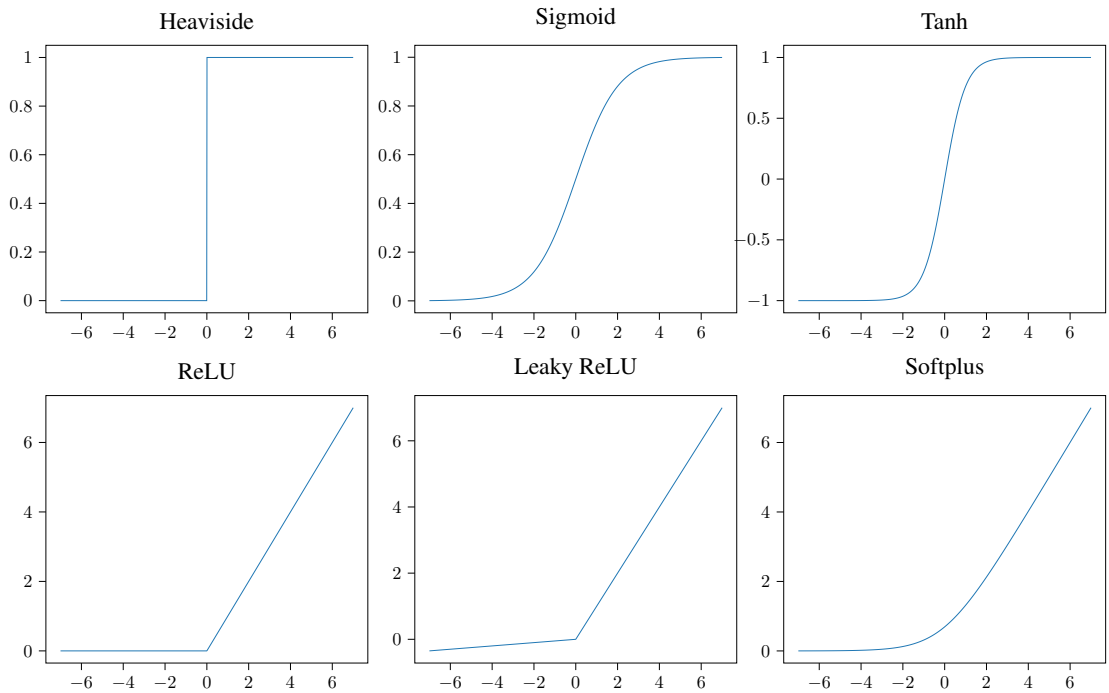


Figure 3.6: Illustration of common activation functions.

3.2.5 Dropout Layers

When a model is trained on a relatively small dataset, it tends to overfit on this training data, resulting in a poor generalization quality on unseen test data. Dropout was introduced for regularization of the model to prevent this overfitting [175]. During training, random parameters of the model are ignored, which approximates training a large number of neural networks with different architectures in parallel. While dropout is commonly included after fully connected layers, there is also work that proposes to apply them after convolutional layers [133]. As an alternative to masking random model parameters, dropout can also be applied to the channels of the feature maps [22].

3.2.6 Downsampling and Upsampling Layers

In many convolutional architectures for image processing, the dimensions of the feature maps S are changed between subsequent layers. An overview of such an architecture is presented in Figure 3.7.

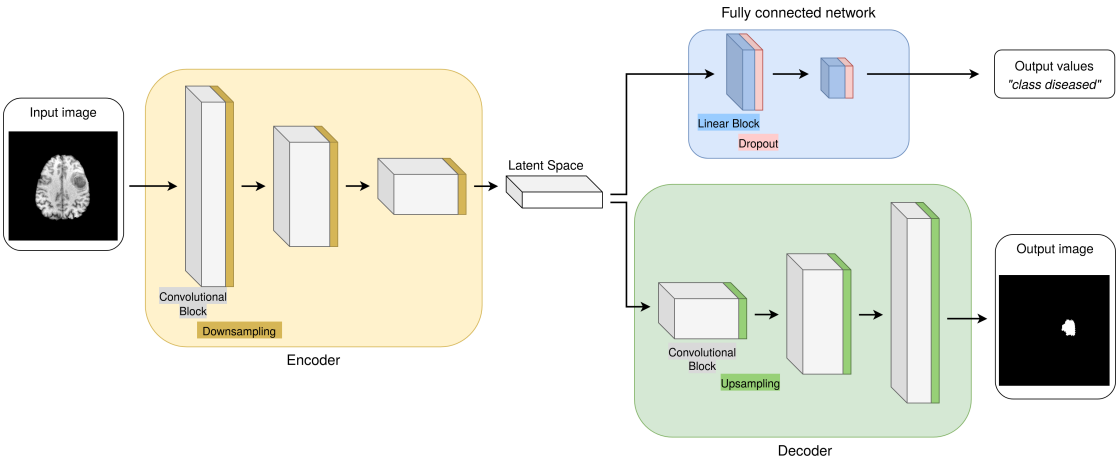


Figure 3.7: Illustration of a typical network architecture for classification, regression, segmentation or reconstruction tasks. Each convolutional or linear block consists of a convolutional or fully connected layer respectively, followed by a normalization layer and an activation function.

First, the image dimensions are reduced to extract lower-level features. At the same time, the number of channels is increased to ensure that the model has enough learnable parameters to extract the features of interest. This architecture is called an *encoder*. Downsampling can be achieved with strided convolutions of stride larger than 1, such that the output image dimensions are smaller than the input image dimensions. Another option is placing pooling layers after the convolutional layers. Popular pooling functions are average or maximum pooling, where multiple input values are summarized to one output value. The output of the encoder is the low-dimensional *latent space* or *feature space*.
For classification or scalar regression tasks, the latent space is passed through a sequence of linear layers, until a scalar prediction is reached. This is illustrated in the upper branch of

Figure 3.7. For segmentation, reconstruction, or image-to-image translation tasks, the encoder is completed with a decoder such that the output image and the input image are of the same size. This is shown in the lower branch of Figure 3.7. The decoder consists of convolutional blocks, followed by upsampling layers. For upsampling, nearest neighbor upsampling or transposed convolutions are regularly used [46, 105].

3.3 Generative Models

The term *generative models* refers to neural networks that learn to generate new data instances. By learning the data distribution of an original dataset $\mathcal{O}$, the models learn to imitate this distribution and generate samples that plausibly could be part of the original dataset $\mathcal{O}$. Deep generative modelling can be divided into variational autoencoders (VAEs), autoregressive models, normalizing flows, energy-based models and generative adversarial nets (GANs) [18].
A VAE [94] is an autoencoder whose latent space distribution is regularised during the training, allowing us to generate new data by sampling from the latent space. A downside of VAEs are the blurry output images due to the training with an MSE loss. Autoregressive models such as PixelRNN [186] or DRAW [56] implicitly define a distribution over a sequence, whereby in each step the next sequence value is predicted given the past values. This can be used for image generation by predicting one pixel value after another. Normalizing flows [145] aim to map a simple distribution to a complex one. This mapping uses invertible functions, resulting in a deterministic transformation. However, they come at a high computational cost. Energy-based models [98] represent probability distributions over data by assigning an unnormalized probability scalar to each input data point. Denoising diffusion models [63, 172] are a subclass of energy-based models, described in detail in Section 3.3.2. GANs are presented in Section 3.3.1.

3.3.1 Generative Adversarial Networks

GANs [55] are based on a game-theory approach and have shown an impressive performance in image generation [20, 85]. Two separate networks, i.e., the generator G_θ with parameters θ and the discriminator D_ϕ with parameters ϕ, play an adversarial game. The goal of G_θ is to generate fake images that could plausibly originate from the original dataset $\mathcal{O}$, whereas the goal of D_ϕ is to distinguish between fake images and real images. In the initial formulation, a generator network consists of a decoder, which generates a fake image out of a random vector $z \sim \mathcal{N}(0, \mathbf{I})$. The generator aims to match the distribution of generated images p_G with the distribution $p_\mathcal{O}$ of the original dataset $\mathcal{O}$, such that the discriminator cannot decide whether a given image is real or fake. The goal is to find the optimal parameter configuration θ^*, ϕ^* to optimize the GAN objective $v(G_\theta, D_\phi)$, given by

$$v(D_\phi, G_\theta) = \mathbb{E}_{x\sim p_\mathcal{O}}[\log D_\phi(x)] + \mathbb{E}_{z\sim\mathcal{N}(0,\mathbf{I})}[\log(1 - D_\phi(G_\theta(z)))]. \tag{3.14}$$

The generator and discriminator are trained iteratively. The generator G_θ aims to minimize the objective function (3.14), while D_ϕ aims to maximize it. The payoff in this zero-sum game for each player is defined as $\pm v(G_\theta, D_\phi)$, where the generator gets the positive payoff and the

discriminator a negative one. During parameter optimization, we use stochastic gradient descent to update the parameters θ of the generator:

$$\nabla_\theta V(G_\theta, D_\phi) = \nabla_\theta \left[\log\left(1 - D_\phi(G_\theta(z))\right)\right], \quad \text{for } z \sim \mathcal{N}(0, \mathbf{I}). \tag{3.15}$$

As the discriminator gets a negative payoff $-v(G_\theta, D_\phi)$, we use gradient ascent to its parameters ϕ of the discriminator:

$$\nabla_\phi V(G_\theta, D_\phi) = \nabla_\phi \left[\log D_\phi(x) + \log(1 - D_\phi(G_\theta(z)))\right], \quad \text{for } z \sim \mathcal{N}(0, \mathbf{I}), x \in \mathcal{O}. \tag{3.16}$$

Ideally, the generator and discriminator converge the Nash equilibrium, i.e., to a stable solution where the players cannot improve their loss objective anymore. This corresponds to a minimization of the Jensen-Shannon divergence between p_G and $p_\mathcal{O}$ [55]. The optimum is reached when p_G gets very close to $p_\mathcal{O}$, and the optimal value for the discriminator reaches 0.5, meaning that the discriminator cannot distinguish between fake and real images any longer.
A major challenge of GANs is the training process due to the adversarial training, where convergence cannot be guaranteed [156]. Moreover, it was shown that GANs do not always have a Nash equilibrium [50]. The training was improved in DCGAN [139]. Implemented changes include replacing fully connected layers with convolutional layers, implementing upsampling convolutions, changing the activation functions, and adding batch normalization to the generator and the discriminator. Mode collapse, non-convergence, and gradient instability remain significant issues despite those improvements.
Wasserstein GANs [7] propose a new loss objective whose mathematical properties improve the gradient stability and prevent a mode collapse. The Wasserstein or Earth Mover's distance between $p_\mathcal{O}$ and p_G is defined by

$$W(p_\mathcal{O}, p_G) = \inf_{\gamma \sim \Pi(p_\mathcal{O}, p_G)} \mathbb{E}_{(x,y)\sim\gamma}[\|x - y\|], \tag{3.17}$$

where $\Pi(p_\mathcal{O}, p_G)$ is the set of all possible joint probability distributions between $p_\mathcal{O}$ and p_G. Using the Kantorovich-Rubinstein duality, the new loss objective aims to minimize a reasonable and efficient approximation $\tilde{W}$ of the Earth-Mover Distance W:

$$\tilde{W}(p_\mathcal{O}, p_G) = \sup_{\|D_\phi\|_L \leq 1} \mathbb{E}_{x\sim p_\mathcal{O}}[D_\phi(x)] - \mathbb{E}_{x\sim p_G}[D_\phi(x)], \tag{3.18}$$

meaning that the discriminator D_ϕ must be 1-Lipschitz continuous. Using these improvements, the GAN objective (3.14) changes to

$$\min_G \max_{\|D_\phi\|_L \leq 1} \mathbb{E}_{\boldsymbol{x}\sim p_\mathcal{O}}[D_\phi(x)] - \mathbb{E}_{z\sim\mathcal{N}(0,\mathbf{I})}[D_\phi(G_\theta(z))]. \tag{3.19}$$

The loss functions for the discriminator $\mathcal{L}_D$ and for the generator $\mathcal{L}_G$ are given by

$$\mathcal{L}_D = W(p_\mathcal{O}, p_G) = \sup_{\|D_\phi\|_L \leq 1} \left(-\mathbb{E}_{x\sim p_\mathcal{O}}[D_\phi(x)] + \mathbb{E}_{z\sim\mathcal{N}(0,\mathbf{I})}[D_\phi(G_\theta(z))]\right), \tag{3.20}$$

$$\mathcal{L}_G = -\mathbb{E}_{z\sim\mathcal{N}(0,\mathbf{I})}[D_\phi(G_\theta(z))]. \tag{3.21}$$

To ensure that the discriminator lies within the space of 1-Lipschitz functions, [58] proposes a gradient penalty loss, changing the loss objective for the discriminator to

$$\mathcal{L}_D = \mathbb{E}_{z\sim\mathcal{N}(0,\mathbf{I})}[D_\phi(G_\theta(z))] - \mathbb{E}_{x\sim p_\mathcal{O}}[D_\phi(x)] + \mathbb{E}_{\hat{x}\sim p_{\hat{x}}}[(\|\nabla_{\hat{x}} D_\phi(\hat{x})\|_2 - 1)^2], \tag{3.22}$$

where the distribution $p_{\hat{x}}$ is sampled uniformly along straight lines between pairs of points sampled from $p_\mathcal{O}$ and p_G.
As proposed in Pix2Pix [74], GANs can also be used for image-to-image translation. This idea was adapted in various methods such as StarGAN [33] or CycleGAN [208]. The architecture of the generator is changed to a U-Net, such that an input image $x \in \mathbb{R}^{h\times w}$ is mapped to a fake image $G(x) \in \mathbb{R}^{h\times w}$, where h and w denote the image dimensions. The discriminator is changed to a patch-discriminator, such that only image patches are considered for the classification into real or fake.

3.3.2 Denoising Diffusion Models

While first presented in [172], diffusion models had a huge success in 2021, when it could be shown that they are able to beat GANs on image synthesis [42]. Denoising diffusion probabilistic models (DDPMs) [63] are based on a parameterized Markov chain. During the diffusion process, for many time steps T, noise is added to an image x until the signal is destroyed. This results in a series of noisy images $\{x_0, x_1, ..., x_T\}$, where the noise level is steadily increased from 0 (no noise) to T (maximum noise). For image generation, we start from noise $x_T \sim \mathcal{N}(0, \mathbf{I})$, and remove small amounts of noise in all time steps $t \in \{T, ..., 1\}$, until we get a fake image x_0. By taking small amounts of Gaussian noise as diffusion, the forward noising process q for a given image $x \in \mathcal{O}$ is defined by the Markov chain

$$q(x_{1:T}|x_0) := \prod_{t=1}^{T} q(x_t|x_{t-1}), \quad \text{with} \quad q(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}), \tag{3.23}$$

with forward process variances $\beta_1, \ldots, \beta_T$, and the identity matrix $\mathbf{I}$. We further define $\alpha_t := 1 - \beta_t$ and $\overline{\alpha}_t := \prod_{s=1}^{t} \alpha_s$. With the reparametrization trick, we can directly write x_t as a function of x_0:

$$x_t = \sqrt{\overline{\alpha}_t}x_0 + \sqrt{1-\overline{\alpha}_t}\epsilon, \quad \text{with } \epsilon \sim \mathcal{N}(0, \mathbf{I}). \tag{3.24}$$

The reverse process is defined by the joint distribution $p_\theta(x_{0:T})$. The starting point is random Gaussian noise $p(x_T) = \mathcal{N}(x_T; 0, \mathbf{I})$. Each step of the reverse process is given by a learned Gaussian transition

$$p_\theta(x_{0:T}) := p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t), \quad \text{with} \quad p_\theta(x_{t-1}|x_t) := \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)), \tag{3.25}$$

where the mean $\mu_\theta(x_t, t)$ and the variance $\Sigma_\theta(x_t, t)$ of p_θ are predicted by the diffusion model. The diffusion model is trained to find the reverse process p_θ that minimizes the variational

upper bound on the negative log likelihood:

$$-\log p_\theta(\mathbf{x}_0) \leq \mathbb{E}_{q(\mathbf{x}_{0:T})}\Big[\log \frac{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}{p_\theta(\mathbf{x}_{0:T})}\Big] =: \mathcal{L}_{vlb} \tag{3.26}$$

As derived in [172], (3.26) can be written in terms of the Kullback-Leibler divergence D_{KL}, meaning that we aim to match $p_\theta(x_{t-1}|x_t)$ and $q(x_{t-1}|x_t, x_0)\,\forall t$:

$$\mathcal{L}_{vlb} = \mathbb{E}_q[\underbrace{D_{\mathrm{KL}}\big(q(x_T|x_0) \parallel p_\theta(x_T)\big)}_{L_T} + \sum_{t=2}^{T} \underbrace{D_{\mathrm{KL}}\big(q(x_{t-1}|x_t, x_0) \parallel p_\theta(x_{t-1}|x_t)\big)}_{L_{t-1}} \underbrace{-\log p_\theta(x_0|x_1)}_{L_0}]. \tag{3.27}$$

L_T is constant and can therefore be ignored during the training. L_0 is approximated with an independent discrete decoder derived from $\mathcal{N}(x_0; \boldsymbol{\mu}_\theta(x_1, 1), \boldsymbol{\Sigma}_\theta(x_1, 1))$, as proposed in [63]. The variance Σ_θ can be fixed to $\Sigma_\theta(x_t, t) = \sigma_t^2 I$, for $\sigma_t^2 = \beta_t$ or $\sigma_t^2 = \tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$, as proposed in [63]. Alternatively, the diffusion model can be trained to predict the interpolation vector v between β_t and $\tilde{\beta}_t$:

$$\Sigma_\theta(x_t, t) = \exp(v \log \beta_t + (1 - v) \log \tilde{\beta}_t). \tag{3.28}$$

By defining $\tilde{\mu}_t(x_t, x_0) := \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t} x_0 + \frac{\sqrt{\bar{\alpha}_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} x_t$ and exploiting the fact that we compare two Gaussian distributions, the loss objective is given by the MSE loss

$$L_{t-1} = \mathbb{E}_{x_0, \epsilon \sim \mathcal{N}(0, \mathbf{I})}\Big[\frac{1}{2\sigma_t^2}\|\tilde{\mu}_t(x_t, x_0) - \mu_\theta(x_t, t)\|^2\Big] + C. \tag{3.29}$$

By reparametrizing

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}\Big(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)\Big) \tag{3.30}$$

for a more stable training, the loss function simplifies to

$$L_{t-1} = \mathbb{E}_{x_0, \epsilon \sim \mathcal{N}(0, \mathbf{I})}\left[\frac{\beta_t^2}{2\sigma_t^2 \alpha_t (1-\bar{\alpha}_t)} \left\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)\right\|^2\right], \quad \text{for } t \in \{2, .., T\}. \tag{3.31}$$

The model follows the architecture of a U-Net with self-attention layers and predicts $\epsilon_\theta(x_t, t)$ from x_t for any step $t \in \{1, ..., T\}$. Information about the time step t is added at every block of the U-Net through a learned embedding. During training, a random time step t is chosen, and the model is updated with the loss objective (3.31).

Sampling Process

During sampling, the goal is to generate a synthetic image out of random noise. This synthetic image should follow the training data distribution. Given x_t and the output of the U-Net model $\epsilon_\theta(x_t, t)$, we can compute the slightly denoised image x_{t-1} using

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\left(\frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t, t)}{\sqrt{\bar{\alpha}_t}}\right) + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2}\epsilon_\theta(x_t, t) + \sigma_t \epsilon, \tag{3.32}$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. During sampling, we start from noise $x_T \sim \mathcal{N}(0, \mathbf{I})$, apply (3.32) for $t \in \{T, ..., 1\}$, until we get a fake image x_0. In Figure 3.8, such a denoising process is presented for a diffusion model trained on the CheXpert dataset [72] of X-ray images of the lungs. This generation process can be conditioned on an input image [32, 154] or on a class label [127]. The advantage of denoising diffusion models is the straightforward training process: Only a U-Net needs to be trained with an MSE loss. This is much more stable than the adversarial training of GANs. A drawback are the long sampling times, since we iteratively need to go through T time steps to generate one image.

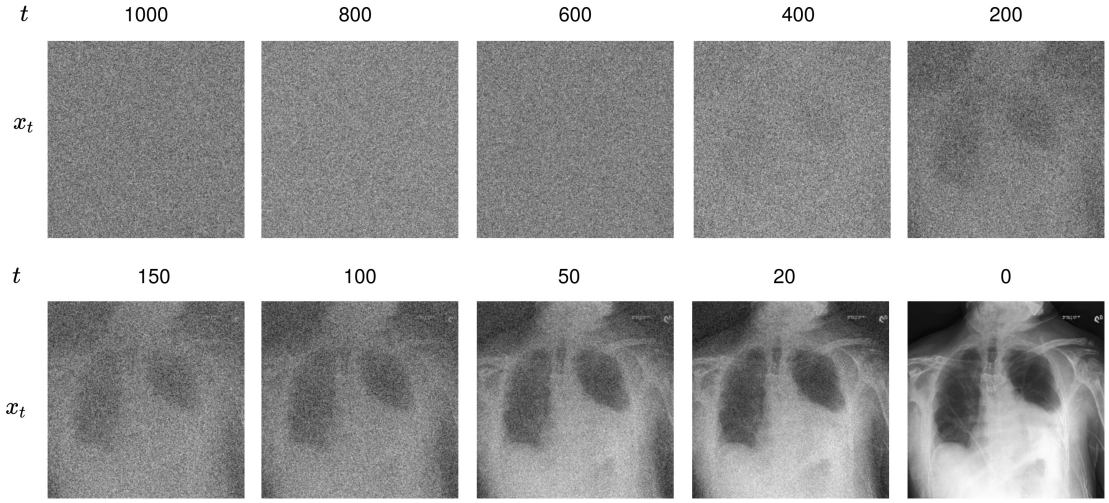


Figure 3.8: Denoising process for the generation of a synthetic image x_0 out of Gaussian noise.

In (3.32) the choice of σ_t defines the type of the generative process [173]. By choosing σ_t accordingly, we can distinguish between the stochastic sampling process described in the DDPM approach, and the deterministic sampling process proposed in the denoising diffusion implicit model (DDIM) formulation [173]. The two approaches are described below.

DDPM Sampling Scheme If we choose $\sigma_t = \sqrt{(1-\bar{\alpha}_{t-1})/(1-\bar{\alpha}_t)}\sqrt{1-\bar{\alpha}_t/\bar{\alpha}_{t-1}}$, we can rewrite(3.32) as

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)\right) + \sigma_t\epsilon, \quad \text{with } \epsilon \sim \mathcal{N}(0, \mathbf{I}). \tag{3.33}$$

We can see in (3.33) that the sampling has a random component ϵ, which leads to a stochastic sampling process. Consequently, by starting from the same initial x_T twice and going through (3.33) for $t \in \{T, ..., 1\}$, we get different output images x_0.

DDIM Sampling Scheme In DDIMs, however, we set $\sigma_t = 0$ in (3.32) , which results in a deterministic sampling process. The loss objective (3.31) for training is still valid. The connection

to ordinary differential equations (ODEs) can be seen when we rewrite (3.32) as

$$\frac{x_{t-1}}{\sqrt{\bar{\alpha}_{t-1}}} = \frac{x_t}{\sqrt{\bar{\alpha}_t}} + \left(\sqrt{\frac{1-\bar{\alpha}_{t-1}}{\bar{\alpha}_{t-1}}} - \sqrt{\frac{1-\bar{\alpha}_t}{\bar{\alpha}_t}}\right)\epsilon_\theta(x_t, t). \tag{3.34}$$

This can be interpreted as the Euler approximation of an ODE. Given infinitely small steps t, the reversed ODE can then be solved with

$$\frac{x_{t+1}}{\sqrt{\bar{\alpha}_{t+1}}} = \frac{x_t}{\sqrt{\bar{\alpha}_t}} + \left(\sqrt{\frac{1-\bar{\alpha}_{t+1}}{\bar{\alpha}_{t+1}}} - \sqrt{\frac{1-\bar{\alpha}_t}{\bar{\alpha}_t}}\right)\epsilon_\theta(x_t, t). \tag{3.35}$$

The encoding scheme presented in (3.35) enables us to go back and forth in the noising and denoising process without losing information. Starting from an original image $x = x_0$ and applying (3.35) for $t \in \{0, ..., T-1\}$, we can obtain encodings x_T of x. In the second step, applying (3.34) on x_T for $t \in \{T, ..., 1\}$ results in the original image x. This combination of encoding and decoding can be used for image interpolation [173]. Furthermore, by applying changes to the training process, DDIMs were adapted image-to-image translation tasks using this iterative deterministic noising and denoising process [103, 136].

Classifier guidance Classifier guidance is introduced in [42]. Let $p_\Phi(y|x_t, t)$ be a classification model trained on the set of noisy images $\{x_0, ..., x_T\}$. Then, the gradient of the classifier $\nabla_{x_t} \log p_\phi(y|x_t, t)$ is used to guide the sampling process towards a desired class label y. In case of the DDPM sampling scheme, x_{t-1} is retrieved by sampling

$$x_{t-1} \sim \mathcal{N}(\mu_\theta(x_t, t) + s\Sigma_\theta(x_t, t)\nabla_{x_t} \log p_\phi(y|x_t), \Sigma_\theta(x_t, t)), \tag{3.36}$$

where the gradient $\nabla_{x_t} \log p_\phi(y|x_t, t)$ is amplified with a constant gradient scale s. In the case of the DDIM sampling scheme, we apply the score-based conditioning trick proposed in [174], and compute an updated epsilon prediction $\hat{\epsilon}$:

$$\hat{\epsilon}(x_t) := \epsilon_\theta(x_t) - s\sqrt{1-\bar{\alpha}_t}\nabla_{x_t} \log p_\phi(y|x_t), \tag{3.37}$$

resulting in an updated sampling scheme

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\left(\frac{x_t - \sqrt{1-\bar{\alpha}_t}\hat{\epsilon}(x_t)}{\sqrt{\bar{\alpha}_t}}\right) + \sqrt{1-\bar{\alpha}_{t-1}}\hat{\epsilon}(x_t). \tag{3.38}$$

An overview for this sampling process using classifier guidance is illustrated in Figure 3.9. Starting from random noise $x_T \sim \mathcal{N}(0, \mathbf{I})$, for every time step $t \in \{T, ..., 1\}$, x_t is passed through the two separate networks of the diffusion model and the classification model. By applying (3.36) or (3.38) to predict x_{t-1}, the image generation is guided towards a desired class y.

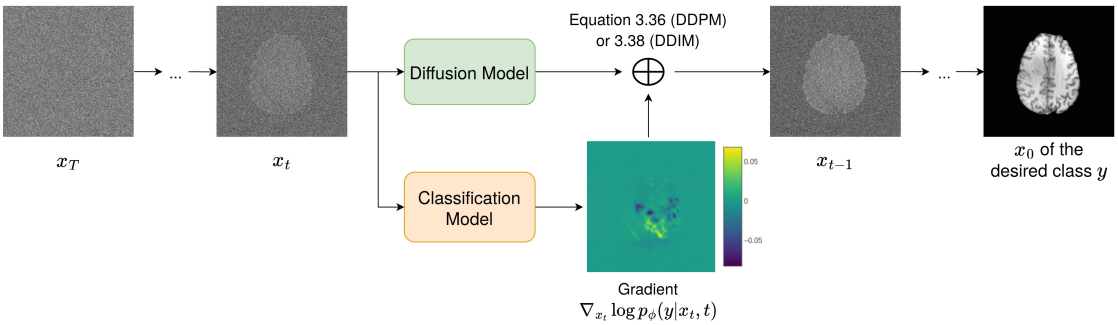


Figure 3.9: Workflow of classifier guidance during the sampling process. In every time step t, the gradient of the classification network is combined with the output of the diffusion model to predict x_{t-1}.

Chapter 4

Deep Learning in Medical Image Analysis

Medical images are an essential part of the clinical diagnostic pipeline. With the increasing number of images, automatic processing of the images is essential in order not to slow down this pipeline. Different automatic image analysis applications are used to support physicians. Image segmentation allows the highlighting of anatomical structures to support their fast identification. Anomaly detection does not highlight specific anatomical structures but parts of the image that differ from the norm. The generated anomaly maps support the detection of pathologies that might not be in focus in the first place. With the increasing number of different imaging modalities, domain adaption has become a fundamental tool for developing novel analysis methods across different acquisition settings.

4.1 Segmentation

Image segmentation is the task of subdividing an input image into regions such that similar parts of the image are grouped in the same segment. This results in a pixel-wise segmentation mask. Let $\mathcal{I}$ be an image with pixel space $\Omega = \{x_{i,j}\}_{i,j=1}^{m,n}$, where $\mathcal{I}$ maps every coordinate in Ω to a pixel value in $\mathbb{R}$:

$$\mathcal{I} : \Omega \to \mathbb{R}. \tag{4.1}$$

Given an image region $R \subseteq \Omega$, a boolean function $b(R, \mathcal{I})$ checks whether the image region $\mathcal{I}(R)$ fulfills some desired segmentation property. The segmentation is then defined by a partition $\Omega = \cup_{i=1}^{k} \Omega_i$, such that

$$\Omega_i \cap \Omega_j = \varnothing \quad \text{for } i \neq j, \tag{4.2}$$

$$b(\Omega_i, \mathcal{I}) = \text{True} \; \forall \, i \in \{1, ..., k\}, \tag{4.3}$$

$$b(\Omega_i \cup \Omega_j, \mathcal{I}) = \text{False} \quad \text{if } i \neq j. \tag{4.4}$$

This task has been applied on a wide range of medical applications for segmenting anatomical structures or lesions [112]. An example is provided in Figure 4.1, where an MR image of the

BRATS2020 dataset is composed of the background segment, the tumor segment, and the segment showing healthy brain tissue.

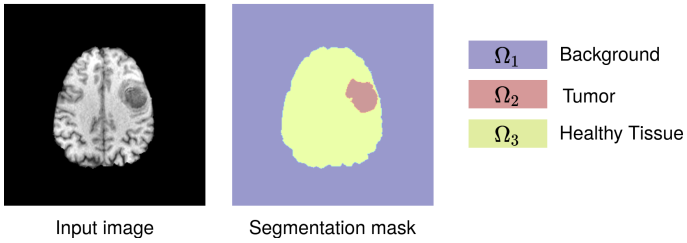


Figure 4.1: Exemplary segmentation mask of a brain MR image. The three segments include the background pixels, the areas showing healthy brain tissue, as well as the pixels belonging to the tumor.

The task of segmentation can be described as a pixel-wise classification. Traditional methods include histogram-based methods [138], thresholding [95], clustering [34], region growing [131], edge detection [157], mathematical morphology [29], atlas-based [80], or graph partitioning methods [47]. Using deep learning, U-Net architectures [147] have shown an outstanding performance in generating the pixel-wise segmentation mask out of an input image [73]. A U-Net consists of an encoder and a decoder that are interconnected with skip connections at various levels. A visualization of such an architecture is given in Figure 4.2. In addition to the low-

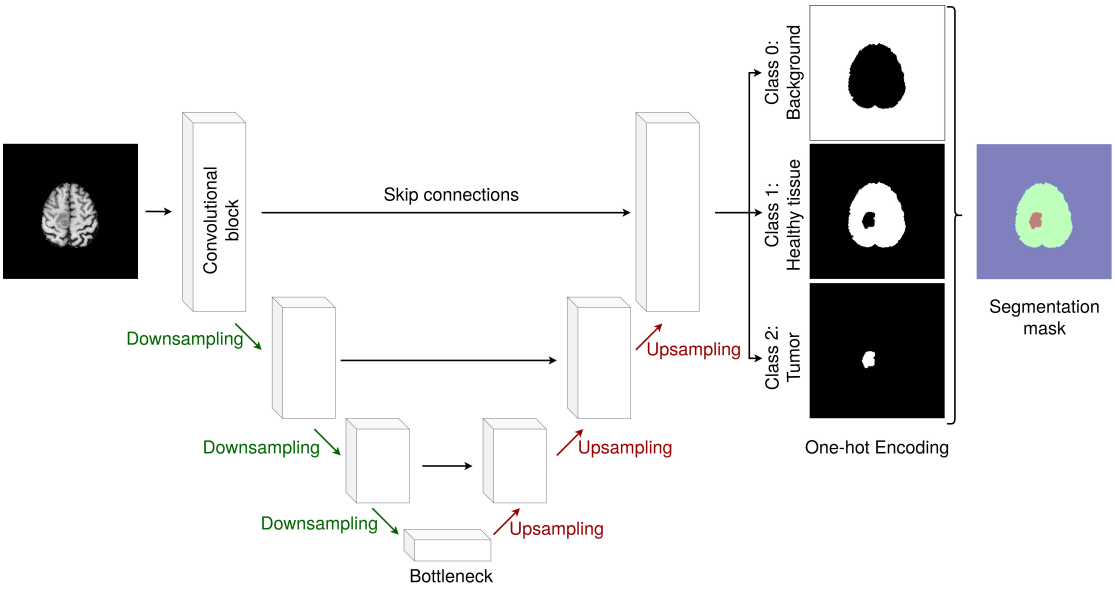


Figure 4.2: U-Net architecture for a segmentation task of a brain MR image. Since there are three classes to segment (background, tumor, and healthy tissue), the output of the model consists of three channels for one-hot encoding of the all classes.

dimensional features extracted by the downsampling process, skip connections add high-level information to the decoder by skipping the low-level feature extraction blocks. In this way, accurate shapes can be reconstructed. The model has multiple output channels for one-hot encoding of the different class labels [142].
State-of-the-art is given by nnU-Net [73], where the model automatically configures itself. This self-configuring optimization for any new task includes preprocessing, network architecture, training, and post-processing. Other segmentation methods use multi-dimensional gated recurrent units [6, 65], which are based on bi-directional recurrent neural networks. Thereby, all spatial dimensions are treated as the temporal sequence, and the segmentation prediction for one voxel is made based on the predictions of the previous voxels. Using a region-based segmentation approach, mask R-CNN [61] is a widely used approach also in the medical field [4, 166]. DeepLab [30] proposes an encoder-decoder architecture based on dilated convolutions and fully connected conditional random fields. This approach was applied to medical tasks such as colorectal polyps or gastric cancer [189, 202]. As an alternative to convolutional neural networks, vision transformers [45] can be adapted for segmentation tasks [176]. The advantage thereby is that global image context is passed through the network at every layer [176].
Popular loss functions for training are the pixel-wise cross-entropy loss or the soft Dice loss described in Section 3.1.1. A significant issue is that many pixel-wise labels are required during training, which is not always possible. To circumvent this issue, semi-supervised segmentation approaches were proposed [100, 110], which opens the possibility of adding unannotated data.
Another issue is that the generated segmentation masks provide no information about the model's decisions. Therefore, methods for pixel-wise uncertainty estimation for the predicted segmentation masks were proposed [79, 122]. One can distinguish between *epistemic* uncertainty, i.e., uncertainty in the model parameters, and *aleatoric* uncertainty, which captures ambiguity and noise in the input data [89]. Common approaches are using the softmax entropy, Monte Carlo dropout during testing, ensembling multiple models, modeling the aleatoric uncertainty using an additional loss term, or using an auxiliary network for uncertainty prediction [79]. We present another approach for interpretable, fully supervised segmentation in Chapter 5.

4.2 Anomaly Detection

Deep learning is a valuable tool for outlier detection, i.e., finding instances that form an exception to a general rule or pattern. In this sense, an outlier - or *anomaly* - is defined as a deviation from the normal behaviour [3]. In real-world applications, anomaly detection algorithms are used for fraud detection, network intrusion detection, or event detection in sensory networks [171]. Traditional techniques include statistical modeling, nearest neighbor methods, clustering, histogram analysis, or principal component analysis [123]. In our medical setting, the normal data is given by a dataset of images of healthy subjects. Images of patients showing a pathology should be identified as outliers. Moreover, this thesis focuses on pixel-level anomaly detection, i.e., highlighting the image regions that show pathological change. We distinguish between two scenarios presented in Figure 4.3.
In scenario A, a network is only trained on the normal control group $\mathcal{H}$. At test time, a new data point is identified as an outlier if it does not follow the expected distribution, i.e., shows

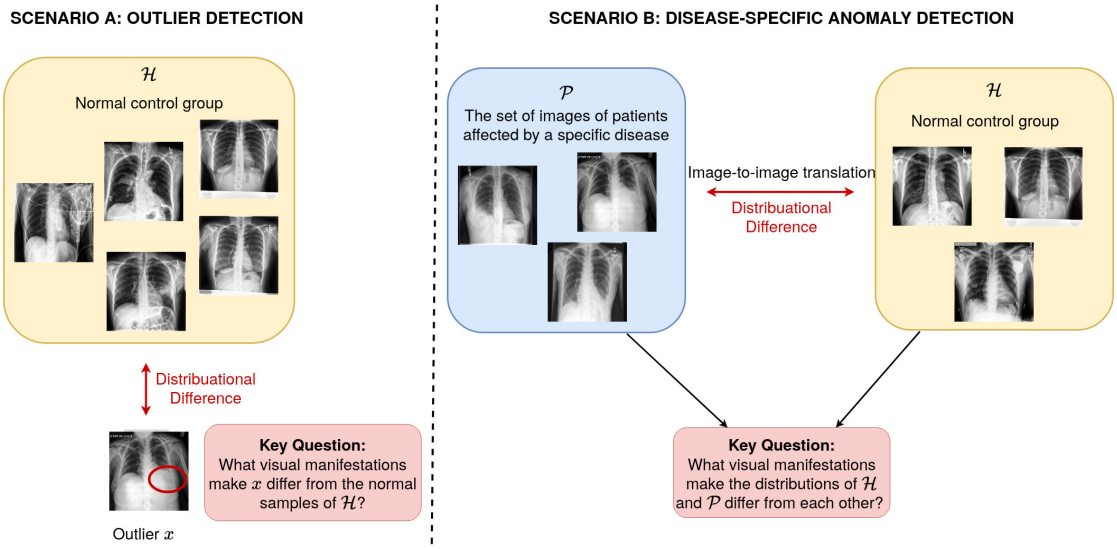


Figure 4.3: In scenario A, we aim to find data instances that do not fit the normal control group $\mathcal{H}$. An image x not matching the distribution of $\mathcal{H}$ is detected as outlier. The image regions that contribute to the abnormality of x are encircled in red. In scenario B, a disease-specific dataset $\mathcal{P}$ is added during training to detect the visual manifestations that make $\mathcal{H}$ and $\mathcal{P}$ differ from each other.

some pathology. One of the main challenges is that samples from the normal control group $\mathcal{H}$ can be very diverse due to numerous subject-specific characteristics and differences in the acquisition settings. The model should have the capacity to represent the diverse nature of $\mathcal{H}$ [51]. Density estimation methods first estimate the probability distribution of the normal images or image features. An unseen image is an outlier if it does not meet the estimated distribution. In one-class classification approaches, the goal is to define a decision boundary of the normal dataset in the feature space. Self-supervised classification approaches aim to learn the visual representation of the input image using an auxiliary task [205], which can be used for outlier detection. Reconstruction-based methods allow pixel-wise comparison between input and output images for pixel-level anomaly detection. As a representative example, we present autoencoders in more detail in Section 4.2.1.

In scenario B, two datasets are at hand: Dataset $\mathcal{H}$ contains images of healthy controls, whereas dataset $\mathcal{P}$ contains images of patients suffering from a specific disease. Using weakly supervised methods, the model learns the difference in distribution between $\mathcal{H}$ and $\mathcal{P}$. Unlike scenario A, however, this anomaly detection focuses on the specific disease present in $\mathcal{P}$ and is not trained to detect any outlier. In Section 4.2.2, we present some weakly supervised GAN approaches for anomaly detection that take advantage of this data setup. Furthermore, a binary classification network can be trained to distinguish between the datasets $\mathcal{H}$ and $\mathcal{P}$. Pixel-level anomaly scores can be gained from gradient-based methods, where backpropagated gradients reveal information about the main difference between $\mathcal{P}$ and $\mathcal{H}$ [205]. This approach is described in Section 4.2.3.

4.2.1 Autoencoders

In scenario A, autoencoders learn to efficiently compress input data to a meaningful low-dimensional representation and reconstruct the input image from this reduced representation. A classical autoencoder architecture is presented in Figure 3.7. An encoder E transforms the input information into a low-dimensional feature representation of the normal data $\mathcal{H}$, and a decoder D reconstructs the input image. Once an outlier is passed through the network, the autoencoder will fail to reconstruct this abnormal input, resulting in a high difference between input and output. An example is given in Figure 4.4, where a convolutional autoencoder is trained on the MNIST dataset [39] of handwritten digit images showing the number 8. The loss is given by the MSE loss between the input and reconstructed images. During evaluation, unseen images are passed through the model. If the input image shows the number 8, i.e., is a sample of the normal dataset, the reconstruction error is low. However, if we pass an image of another number, e.g., a 5, through the model, the autoencoder fails to create a good reconstruction, resulting in a high error. This error can be defined as the pixel-level or image-level anomaly score. A drawback of autoencoders is the blurry output images, which result in inaccurate pixel-wise error maps.

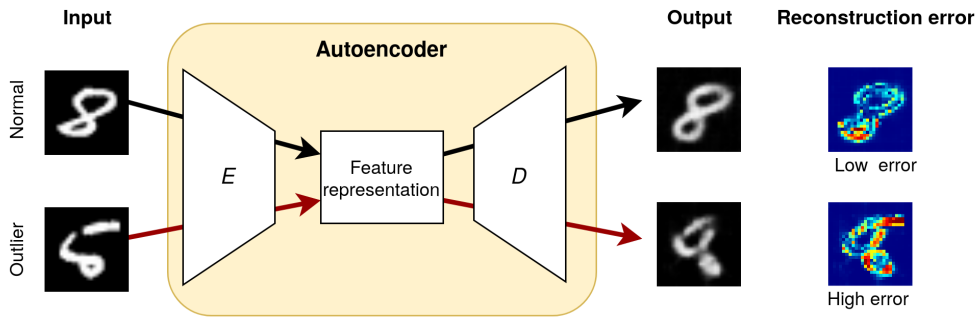


Figure 4.4: Exemplary autoencoder for anomaly detection. The autoencoder is trained on the MNIST dataset on images showing the number 8. During evaluation, the processing of an image showing the number 5 leads to a high reconstruction error, indicating an outlier.

Variational Autoencoders (VAEs) [94] differ from autoencoders in their loss objective and have been applied on a wide range of medical anomaly detection tasks [108, 113, 209]. Instead of encoding the latent space representation of an input x, the encoder learns a latent distribution $\mathcal{N}(\mu, \sigma)$. The sampled latent space representation $z \sim \mathcal{N}(\mu, \sigma)$ is then passed through the decoder. The reconstruction loss is amended with the Kullback-Leibler divergence D_{KL} between the sampled distribution $\mathcal{N}(\mu, \sigma)$ and the target latent distribution $\mathcal{N}(0, \mathbf{I})$,

$$\mathcal{L}_{VAE}(x) = \underbrace{\mathbb{E}_{z\sim\mathcal{N}(\mu,\sigma)}\left[\|x - D(E(z))\|_2\right]}_{\text{reconstruction loss}} + \underbrace{D_{KL}\big(\mathcal{N}(\mu, \sigma) \parallel \mathcal{N}(0, \mathbf{I})\big)}_{\text{KL divergence}} . \tag{4.5}$$

During the evaluation, this class of models can be used as generative models by sampling $z \sim \mathcal{N}(\mu, \sigma)$ and retrieving an output image $D(z)$. For anomaly detection, the Kullback-Leibler divergence D_{KL} can be taken as an anomaly score for an unseen data point [108, 187].

4.2.2 Anomaly Detection with GANs

In their initial formulation presented in Section 3.3.1, GANs are purely generative models with a strong ability to model the training data distribution. If the generator follows an encoder-decoder-architecture that takes images as input, they can be adapted to anomaly detection tasks [43, 161, 201]. In scenario A, with the adversarial training between generator and discriminator, the encoder of the generator learns a latent space representation of the normal data $\mathcal{H}$. AnoGAN [161], and GANomaly [1] use this training scheme to generate pixel-wise anomaly maps based on the reconstruction error. In a data setup like in scenario B, we can perform image-to-image translation between the two sets $\mathcal{H}$ and $\mathcal{P}$ using GANs. Inspired by CycleGAN [208], cycle-consistent translation from normal to anomalous data and vice versa can be used for anomaly detection [15, 155]. Since the anomaly map is defined by the difference between the input image of the set $\mathcal{P}$ and its translation to the set $\mathcal{H}$, it is crucial that only pathological regions are changed during image-to-image translation. Another GAN implementation that needs data of both $\mathcal{H}$ and $\mathcal{P}$ is VAGAN [14], where an additive map is learned to translate anomalous images to images of healthy subjects. The additive map was used to highlight anomalous changes in MR images of the brain of patients who have Alzheimer's disease. In Figure 4.5, the workflows of CycleGAN and VAGAN are visualized.
Similarly, Fixed-Point-GAN [168] builds on CycleGAN also relies on generating an additive map for the input image and proposes additional identity-preserving constraints. With this approach, brain lesions and pulmonary embolisms could be localized. PathoGAN [5] adapts the cycle-consistent image-to-image translation of CycleGAN. For MR images of the brain, an inpainting of healthy-looking tissue is generated in image regions showing a tumor. Thereby, a segmentation mask highlighting the anomalous changes is implicitly learned during the training process.
In Chapter 6, we present another approach for scenario B based on CycleGAN. In contrast to VAGAN or Fixed-Point-GAN, our method is not restricted to the generation of an additive map and has therefore the flexibility to change various image features.

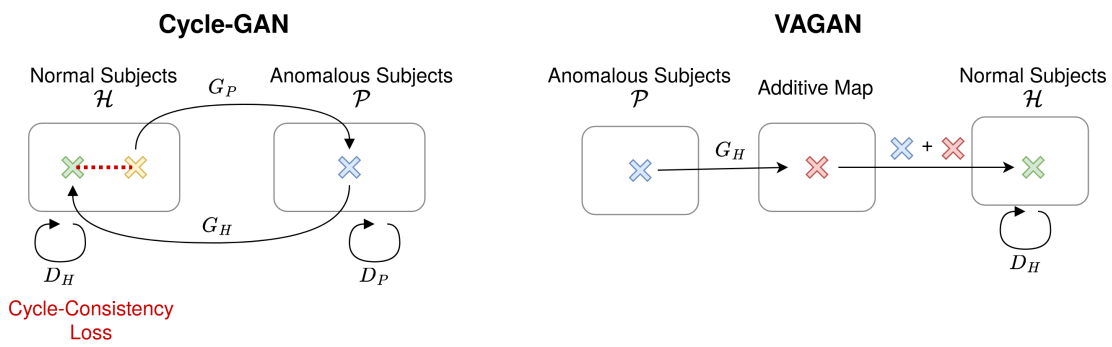


Figure 4.5: Overview of CycleGAN and VAGAN approaches. The generator G_H aims to generate images of normal subjects, whereas the generator G_P aims to generate images of anomalous subjects. The discriminators D_H and D_P distinguish between fake and real images of the sets $\mathcal{H}$ and $\mathcal{P}$ respectively.

4.2.3 Gradient-based Anomaly Detection

In scenario B, the gradients of a classification model between $\mathcal{H}$ and $\mathcal{P}$ can be used for anomaly detection. Saliency maps [169] take the pixel-wise gradient of the prediction with respect to the input image. These saliency maps highlight pixels that contribute to the model's decision and can be interpreted as a pixel-wise anomaly map [8]. Class activation maps [207] were proposed for the interpretability of CNNs. At a given layer k of this network, a global average pooling layer is introduced to compute the importance w_i of the feature maps $S_{i,k}$ at that layer. The sum of the feature maps $S_{i,k}$, weighted by w_i, defines the class activation map C_k:

$$C_k = \sum_i S_{i,k} w_i. \tag{4.6}$$

By upsampling C_k to the size of the input image, the image regions that contribute to the classification of the input image are highlighted. Since the upsampling results in blurry heatmaps, Grad-CAM [163] uses higher-level feature maps and proposes the sum over the gradient of the classifier with respect to the feature maps $S_{i,k}$ as the weight w_i. This opens the possibility of visualizing the activation maps at different feature levels k of the classifier. To illustrate this approach, we train a binary classification network following the VGG16 architecture [170] on the BRATS2020 dataset to distinguish between slices containing a tumor and slices without tumor. In Figure 4.6 we present the results for the class activation map and the saliency map for a slice showing a tumor. The regions that lead to the classification score "diseased" are highlighted in the class activation map. We observe that the region showing the tumor is covered. However, due to the strong upsampling, the heatmap is blurry. The saliency map is more detailed but does not cover the whole tumor. Despite these downsides, those approaches represent an essential building block of more advanced methods, for example, in the gradient guidance of diffusion models presented in Chapter 7.

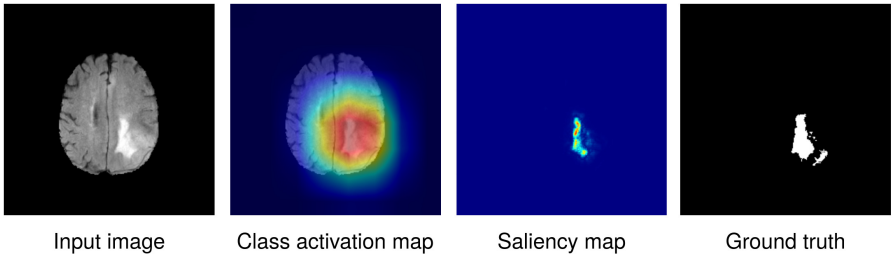


Figure 4.6: Visualization of the class activation map and saliency map using a binary classification model between slices with or without tumor on the BRATS2020 dataset.

4.3 Domain Adaptation

In general machine learning algorithms, it is assumed that the training data follows an underlying distribution, which is the same for the test data. However, in many real-world scenarios, the test data may vary in numerous aspects. Examples include the exact acquisition settings, outer

influence, or differences in the subjects such as an age shift in the population. Therefore, a model optimized on the training set loses some of its performance if applied to a test set that follows a related but different distribution. In the scope of domain adaptation, a model learns a main task from a *source domain*. The goal is to develop methods that perform well on a different *target domain* by overcoming the distribution shift between the source and the target domain. Thereby, the generalization quality of the model is improved. In Figure 4.7, such a domain shift between source and target domain is visualized for the main task of binary classification between squares and circles.

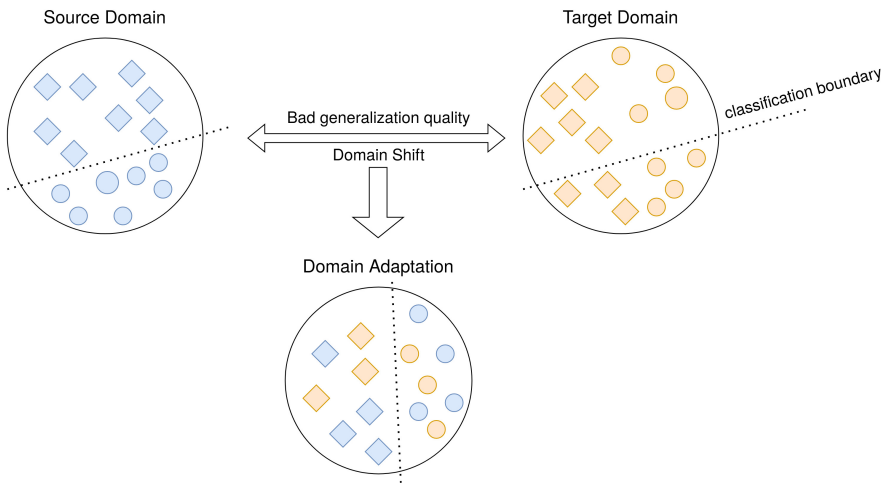


Figure 4.7: The problem of domain adaptation tackles the issue of a distribution shift between the source dataset and the target dataset. The goal is a generalizable model that has a good performance regarding the main task on both domains.

We define the source domain $\mathcal{D}_s = \{x_{i,s}, y_{i,s} | i \in \{1, ..., n\}\}$, where $x_{i,s}$ is the input data and $y_{i,s}$ the corresponding label, which follow a source domain distribution $p_s(x, y)$. The target domain is defined by $\mathcal{D}_t = \{x_{i,t}, y_{i,t} | i \in \{1, ..., m\}\}$, following a target domain distribution $p_t(x, y)$. Generally, we can distinguish between supervised, semi-supervised and unsupervised domain adaptation. While the target labels $y_{i,t}$ are known during training in supervised domain adaptation (SDA), only a few target samples are labeled in semi-supervised domain adaptation. Unsupervised domain adaptation (UDA) describes the field where no labeled data from the target domain is available for training. Furthermore, domain adaptation can be either homogeneous, i.e., the input feature spaces are the same across the domains, or heterogeneous, i.e., the feature spaces and their dimensionalities may differ [190].
In medical applications, domain adaptation is an inevitable step to achieving consistency over different datasets. Since data may be acquired at multiple sites with different acquisition settings and patients, it is challenging to collect large homogeneous datasets to train deep learning models [57]. A first important application field is histopathological imaging, where the images must be stained before image analysis tasks are applied. As samples may have been prepared at different laboratories, this results in heterogeneous staining characteristics because of slight differences in incubation times and the protocol. The variations in stain color considerably reduce

the performance of deep learning models [25, 144]. Other approaches focus on cross-modality domain adaptation, e.g., learning a task on CT images and applying it to MR images as the target domain [28, 77, 134].
Regarding MRI, inhomogeneous datasets due to the scanner bias described in Section 2.1.2 are likely to occur. Therefore, MR harmonization is an important topic and has been tackled in various ways [48, 135, 199]. Several SDA approaches have been proposed for MR image classification or segmentation [66, 88, 184] by fine-tuning the pre-trained model parameters on the target dataset, or with multi-task learning [111]. Some UDA approaches show that extensive data augmentation of MR can improve segmentation results [17, 118, 128, 206].
Generally, the domain adaptation problem can be addressed with divergence, adversarial, or reconstruction-based methods.

Divergence-based methods Divergence-based methods [82, 203] aim to minimize some divergence measure between the source and the target domain distributions p_s and p_t, and thereby bridge the domain shift. To reduce the domain discrepancy, contrastive loss terms are applied to push images with the same task-related characteristics close to each other while placing images with different characteristics far from each other in the feature space. Given an input image x, we consider a similar image x_p and an image x_n that differs from x in any aspect related to the main task. We denote x_p as a *positive sample* and x_n as a *negative sample*. The learned feature representation $f(x)$ should be close to $f(x_p)$ and far from $f(x_n)$. The initial contrastive loss objective [59] is given by

$$\mathcal{L}_{con}(x, z) = \mathbb{1}_{z=x_p}\|f(x) - f(z)\|_2 + \mathbb{1}_{z=x_n} \max(0, m - \|f(x) - f(z)\|_2), \tag{4.7}$$

where m denotes a margin that defines a radius around $f(x)$ in the feature space. This idea was extended to the triplet loss [162], where the image x as well as a positive sample x_p and a negative sample x_n are taken into account:

$$\mathcal{L}_{triplet}(x, x_p, x_n) = \max\big(\|f(x) - f(x_p)\|_2 - \|f(x) - f(x_n)\|_2 + m, 0\big), \tag{4.8}$$

In state-of-the-art domain adaptation methods, the same idea of clustering the feature space is applied, but more elaborate loss terms are used, such as the maximum mean discrepancy (MMD) [82], Kullback-Leibler divergence [126], cosine similarity [31] or the contrastive domain discrepancy (CDD) [49]. We adapt this idea of using contrastive loss terms to align the feature space in Chapter 9.

Adversarial-based methods In adversarial-based methods [53, 81, 181], a model provides feature representations of the input images and is trained to solve a given main task. An additional discriminator network is trained to distinguish between the source and the target domain based on those feature representations. Like in GANs, iterative adversarial training is performed until the discriminator can no longer distinguish between the domains. Consequently, domain-invariant features are extracted by the model. This is useful to train an MR classification, segmentation, or regression task while training the discriminator network to distinguish between the MR scanners. With the adversarial training, MR harmonization can be achieved [44, 130, 159].

Reconstruction-based methods Reconstruction-based methods [67, 204] focus on image-to-image translation between the domains. Images of the target domain are translated to the source domain before a specific task is applied. For this approach, autoencoders or GANs can be used. The challenges are that the generated images should be of high quality and correspond to the input image without adding or losing information.
If paired data is available, pairwise image-to-image translation methods can be taken into account. Deepharmony [41] uses paired data to change the contrast of MR images from one scanner to another with a modified U-Net. This approach is illustrated in Figure 4.8, where the task is to translate 1.5T MR images to 3T MR images. The output shows the desired increase in contrast but also a smoothing of the image. However, paired datasets are not often available for training. GANs [68, 99, 158] follow a similar idea and aim to generate new images to overcome the domain shift without relying on paired data.

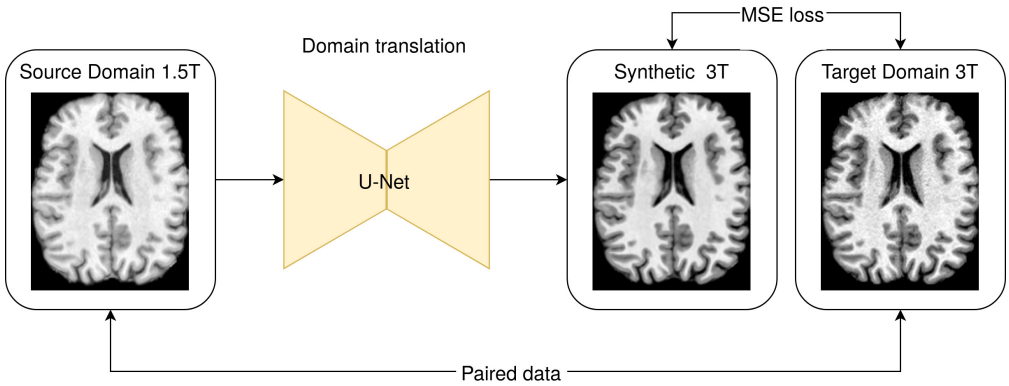


Figure 4.8: Domain translation from the source domain to the target domain with Deepharmony. In this setup, paired data is required.

Chapter 5

Diffusion Models for Implicit Image Segmentation Ensembles

In the following paper, we present a novel approach for fully supervised image segmentation. The generation of the segmentation mask is based on denoising diffusion probabilistic models and takes advantage of their stochastic sampling process. In this way, an ensemble of segmentation masks can be generated for every input image. Consequently, taking the mean and the variance map over the ensemble of segmentation mask, we improve the segmentation performance and automatically generate a pixel-wise uncertainty evaluation. This approach is evaluated on the BRATS2020 dataset for brain tumor segmentation.
The manuscript was written in joint first authorship with Robin Sandkühler, who had the initial idea for this approach. The workload was divided such that the development of the method was a shared contribution. Julia Wolleb performed the code implementation and the experiments. The initial draft was written by Julia Wolleb and revised by Robin Sandkühler.

Publication. The following paper was accecpted at the conference *Medical Imaging with Deep Learning* (MIDL), July 2022, Zurich, Switzerland. The manuscript was published as part of the conference proceedings [197]. The code for this framework is open-source[1].

[1] https://gitlab.com/cian.unibas.ch/Diffusion-based-Segmentation

Diffusion Models for Implicit Image Segmentation Ensembles

Julia Wolleb* JULIA.WOLLEB@UNIBAS.CH
Robin Sandkühler* ROBIN.SANDKUEHLER@UNIBAS.CH
Florentin Bieder FLORENTIN.BIEDER@UNIBAS.CH
Philippe Valmaggia PHILIPPE.VALMAGGIA@UNIBAS.CH
Philippe C. Cattin PHILIPPE.CATTIN@UNIBAS.CH
Department of Biomedical Engineering, University of Basel, Allschwil, Switzerland

Abstract

Diffusion models have shown impressive performance for generative modelling of images. In this paper, we present a novel semantic segmentation method based on diffusion models. By modifying the training and sampling scheme, we show that diffusion models can perform lesion segmentation of medical images. To generate an image-specific segmentation, we train the model on the ground truth segmentation, and use the image as a prior during training and in every step during the sampling process. With the given stochastic sampling process, we can generate a distribution of segmentation masks. This property allows us to compute pixel-wise uncertainty maps of the segmentation, and allows an implicit ensemble of segmentations that increases the segmentation performance. We evaluate our method on the BRATS2020 dataset for brain tumor segmentation. Compared to state-of-the-art segmentation models, our approach yields good segmentation results and, additionally, detailed uncertainty maps.

Keywords: Diffusion models, segmentation, uncertainty estimation

1. Introduction

Semantic segmentation is an important and well-explored area in medical image analysis (Rizwan I Haque and Neubert, 2020). The automated segmentation of lesions in medical images with machine learning has shown good performances (Isensee et al., 2021) and is ready for clinical application to support diagnosis (Sharrock et al., 2021). In medical applications, it is of high interest to measure the uncertainty of a given prediction, especially when used for further treatments like radiation therapy.

In this work, we focus on the BRATS2020 brain tumor segmentation challenge (Menze et al., 2014; Bakas et al., 2017, 2018). This dataset provides four different MR sequences for each patient (namely T1-weighted, T2-weighted, FLAIR and T1-weighted with contrast enhancement), as well as the pixel-wise ground truth segmentation. An exemplary image can be found in Appendix A.

We propose a novel segmentation method based on a Denoising Diffusion Probabilistic Model (DDPM) (Ho et al., 2020), which can provide uncertainty maps of the produced segmentation mask. An overview of the workflow for an image of the BRATS2020 dataset is shown in Figure 1. We train a DDPM on the segmentation masks and add the original brain MR image as an image prior to induce the anatomical information. As sampling with

* Contributed equally

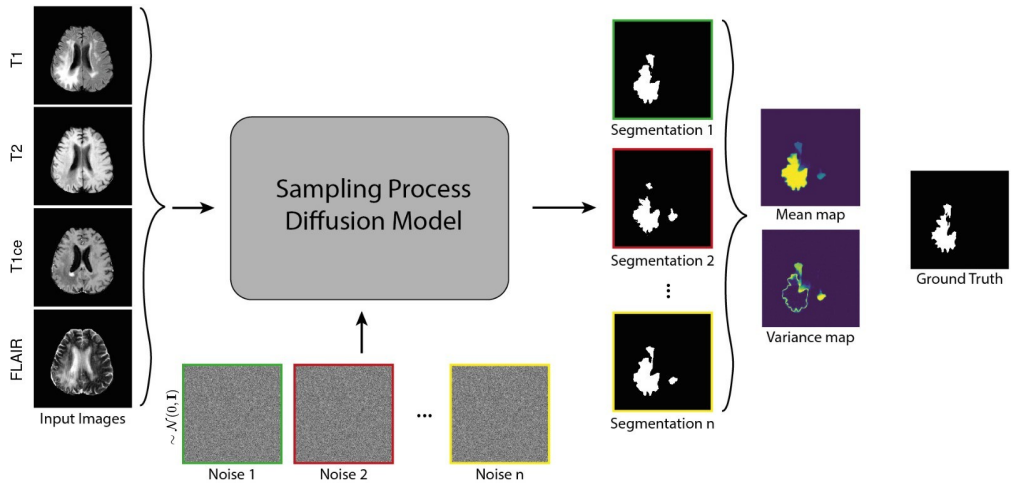


Figure 1: Workflow for the implicit generation of segmentation ensembles and uncertainty maps with diffusion models. The input image consists of four different MR sequences. Going n times through the sampling process of the diffusion model with different Gaussian noise, n different segmentation masks are generated.

DDPMs has a stochastic element in each sampling step, we can generate many different segmentation masks for the same input image and the same pretrained model. This ensemble of segmentations allows us to compute the pixel-wise variance maps, which visualizes the uncertainty of the generated segmentation. Moreover, the ensembling of the segmentations in a mean map boosts the segmentation performance.
We compare ourselves against state-of-the-art segmentation algorithms, and visually compare our variance map against common uncertainty maps. The code is publicly available at `https://gitlab.com/cian.unibas.ch/Diffusion-based-Segmentation`.

Related Work In medical image segmentation, a common method is the application of a U-Net (Ronneberger et al., 2015) or SegNet (Badrinarayanan et al., 2017) to predict the segmentation mask for every input image. This approach was successfully applied for many different tasks (Habijan et al., 2019; Kumar et al., 2019; Xiao et al., 2020). The state of the art is given by nnU-Nets (Isensee et al., 2021), where the best architecture and hyperparameters are automatically chosen for every specific dataset.
Uncertainty quantification is of high interest in deep learning research (Abdar et al., 2021), which is often done using Bayesian neural networks (Kendall et al., 2017; Mitros and Mac Namee, 2019; Gal and Ghahramani, 2016). We can differentiate between epistemic uncertainty, which refers to uncertainty in the model parameters, and aleatoric uncertainty, which refers to uncertainty in the data. As stated in (Kendall and Gal, 2017), the epistemic uncertainty of a segmentation model can be approximated with Monte Carlo Dropout, whereas the aleatoric uncertainty can be modeled with Maximum-A-Posteriori inference. Those methods were also applied on various medical tasks (Wang et al., 2019; Nair et al., 2020; DeVries and Taylor, 2018), including brain tumor segmentation (Sagar, 2020; Jungo

and Reyes, 2019; Mehta et al., 2020). Other approaches presented stochastic segmentation networks to model aleatoric uncertainty (Monteiro et al., 2020), or proposed a probabilistic U-Net to learn a distribution over segmentations (Kohl et al., 2018, 2019).
During the last year, DDPMs have gained a lot of attention due to their astonishing performance in image generation (Dhariwal and Nichol, 2021). Images are generated by sampling from Gaussian noise. This sampling scheme follows a stochastic process, and therefore sampling from the same noisy image does not result in the same output image. A different sampling scheme was introduced by Denoising Diffusion Implicit Models (DDIM)(Song et al., 2020), where sampling is deterministic and can be done by skipping multiple steps. Moreover, meaningful interpolation between images can be achieved. DDPM was further improved by (Nichol and Dhariwal, 2021) and (Dhariwal and Nichol, 2021), where changes in the loss objective, architecture improvements, and classifier guidance during sampling improved the output image quality.
While some new work applies diffusion models on tasks such as image-to-image translation (Sasaki et al., 2021), style transfer (Choi et al., 2021), or inpainting tasks (Saharia et al., 2021), so far there is only very little work about semantic segmentation. Recently, one approach to perform semantic segmentation with a diffusion model was proposed by (Baranchuk et al., 2022). A DDPM is trained to reconstruct the image that should be segmented. Then, a multilayer perceptron for classification is applied on the features of the model, which results in a segmentation mask for the original image. In contrast to this method, we train a DDPM directly to generate the segmentation mask. Simultaneously and independent from us, (Amit et al., 2021) developed an image segmentation method similar to ours. However, they use a separate encoder for the image and the segmentation. Training a larger model may be difficult for medical image analysis due to possible large input images such as 3D data. Our method uses only one encoder to encode the image information and the segmentation mask.

2. Method

The goal is to train a DDPM to generate segmentation masks. We follow the idea and implementation proposed in (Nichol and Dhariwal, 2021). The core idea of diffusion models is that for many timesteps T, noise is added to an image x. This results in a series of noisy images $x_0, x_1, ..., x_T$, where the noise level is steadily increased from 0 (no noise) to T (maximum noise). The model follows the architecture of a U-Net and predicts x_{t-1} from x_t for any step $t \in \{1, ..., T\}$. During training, we know the ground truth for x_{t-1}, and the model is trained with an MSE loss. During sampling, we start from noise $x_T \sim \mathcal{N}(0, \mathbf{I})$, sample for T steps, until we get a fake image x_0.
The complete derivations of the formulas below can be found in (Ho et al., 2020; Nichol and Dhariwal, 2021). The main components of diffusion models are the forward noising process q and the reverse denoising process p. Following (Ho et al., 2020), the forward noising process q for a given image x at step t is given by

$$q(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}), \tag{1}$$

where $\mathbf{I}$ denotes the identity matrix and $\beta_1, ..., \beta_T$ are the forward process variances. The idea is that in every step, a small amount of Gaussian noise is added to the image. Doing

this for t steps, we can write

$$q(x_t|x_0) := \mathcal{N}(x_t; \sqrt{\overline{\alpha}_t}x_0, (1-\overline{\alpha}_t)\mathbf{I}), \tag{2}$$

with $\alpha_t := 1 - \beta_t$ and $\overline{\alpha}_t := \prod_{s=1}^{t} \alpha_s$. With the reparametrization trick, we can directly write x_t as a function of x_0:

$$x_t = \sqrt{\overline{\alpha}_t}x_0 + \sqrt{1-\overline{\alpha}_t}\epsilon, \quad \text{with } \epsilon \sim \mathcal{N}(0, \mathbf{I}). \tag{3}$$

The reverse process p_θ is learned by the model parameters θ and is given by

$$p_\theta(x_{t-1}|x_t) := \mathcal{N}\big(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)\big). \tag{4}$$

As shown in (Ho et al., 2020), we can then predict x_{t-1} from x_t with

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}\epsilon_\theta(x_t, t)\right) + \sigma_t \mathbf{z}, \quad \text{with } \mathbf{z} \sim \mathcal{N}(0, \mathbf{I}), \tag{5}$$

where σ_t denotes the variance scheme that can be learned by the model, as proposed in (Nichol and Dhariwal, 2021). We can see in Equation 5 that sampling has a random component $\mathbf{z}$, which leads to a stochastic sampling process. Note that ϵ_θ is the U-Net we train, with input $x_t = \sqrt{\overline{\alpha}_t}x_0 + \sqrt{1-\overline{\alpha}_t}\epsilon$. The noise scheme $\epsilon_\theta(x_t, t)$ that will be subtracted from x_t during sampling according to Equation 5 has to be learned by the model. This U-Net is trained with the loss objectives given in (Nichol and Dhariwal, 2021).
We now modify this idea to use diffusion models for semantic segmentation. A visualization of the workflow is given in Figure 2 for the task of brain tumor segmentation.

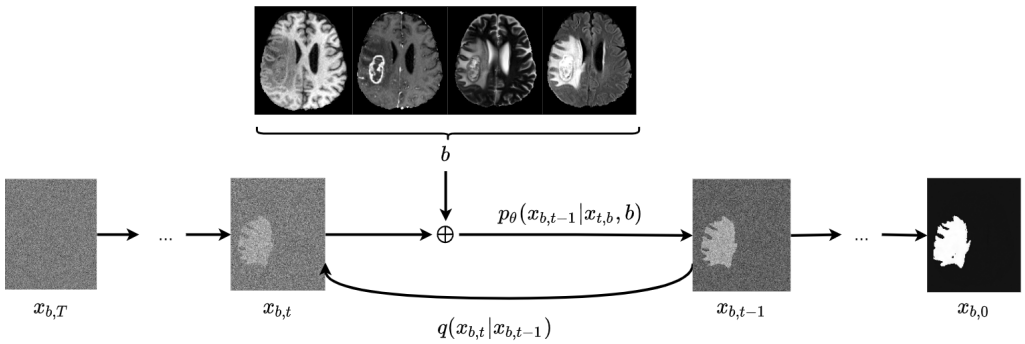


Figure 2: The training and sampling procedure of our method. In every step t, the anatomical information is induced by concatenating the brain MR images b to the noisy segmentation mask $x_{b,t}$.

Let b be the given brain MR image of dimension (c, h, w), where c denotes the number of channels, and (h, w) denote the image height and image width. The ground truth segmentation of the tumor for the input image b is denoted as x_b, and is of dimension $(1, h, w)$. We

train a DDPM for the generation of segmentation masks. In the classical DDPM approach, x_b would be the only input we need for training, which would result in an arbitrary segmentation mask x_0 when we sample from noise during inference. In contrast to that, the goal in our proposed method is not to generate *any* segmentation mask, but we want a meaningful segmentation mask $x_{b,0}$ for a given image b. To achieve this, we add additional channels to the input: We induce the anatomical information present in b by adding it as an image prior to x_b. We do this by concatenating b and x_b, and define $X := b \oplus x_b$. Consequently, X has dimension $(c+1, h, w)$.
During the noising process q, we only add noise to the ground truth segmentation x_b:

$$x_{b,t} = \sqrt{\overline{\alpha}_t} x_b + \sqrt{1-\overline{\alpha}_t}\epsilon, \quad \text{with } \epsilon \sim \mathcal{N}(0, \mathbf{I}), \tag{6}$$

and we define $X_t := b \oplus x_{b,t}$. Equation 5 is then altered to

$$x_{b,t-1} = \frac{1}{\sqrt{\alpha_t}} \left(x_{b,t} - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}} \epsilon_\theta(X_t, t) \right) + \sigma_t \mathbf{z}, \quad \text{with } \mathbf{z} \sim \mathcal{N}(0, \mathbf{I}) \tag{7}$$

and results in a slightly denoised $x_{b,t-1}$ of dimension $(1, h, w)$. During inference, we follow the procedure presented in Algorithm 1, which is a stochastic process. Therefore, sampling twice for the same brain MR image b does not result in the same segmentation mask prediction $x_{b,0}$. Exploiting this property, we can implicitly generate an ensemble of segmentation masks without having to train a new model. This ensemble can then be used to boost the segmentation performance.

Algorithm 1: Sampling Procedure

Input: b, the original brain MRI
Output: $x_{b,0}$, the predicted segmentation mask
sample $x_{b,T} \sim N(0, \mathbf{I})$;
for $t \leftarrow T$ **to** 1 **do**
  $X_t \leftarrow b \oplus x_{b,t}$;
  $x_{b,t-1} \leftarrow \frac{1}{\sqrt{\alpha_t}} \left(x_{b,t} - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}} \epsilon_\theta(X_t, t) \right) + \sigma_t \mathbf{z}, \quad \text{with } \mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$;
end

3. Dataset and Training Details

We evaluate our method on the BRATS2020 dataset. As described in Section 1, images of four different MR sequences are provided for each patient, which are stacked to 4 channels. We slice the 3D MR scans in axial slices. Since tumors rarely occur on the upper or lower part of the brain, we exclude the lowest 80 slices and the uppermost 26 slices. For intensity normalization, we cut the top and bottom one percentile of the pixel intensities. We crop the images to a size of (4, 224, 224). The provided ground truth labels contain four classes, which are background, GD-enhancing tumor, the peritumoral edema, and the necrotic and non-enhancing tumor core. We merge the three different tumor classes into one class and therefore define the segmentation problem as a pixel-wise binary classification. Our training

set includes 16,298 images originating from 332 patients, and the test set comprises 1,082 images with non-empty ground truth segmentations, originating from 37 patients. No data augmentation is applied.
The hyperparameters for our DDPM models are described in the appendix of (Nichol and Dhariwal, 2021). We choose a linear noise schedule for $T = 1000$ steps. The model is trained with the hybrid loss objective, with a learning rate of 10^{-4} for the Adam optimizer, and a batch size of 10. The number of channels in the first layer is chosen as 128, and we use one attention head at resolution 16. We train the model for 60,000 iterations on an NVIDIA Quadro RTX 6000 GPU, which takes around one day. The training details for the comparing methods can be found in Appendix B.

4. Results and Discussion

During evaluation, we take an image b from the test set, follow Algorithm 1 and produce a segmentation mask. This mask is thresholded at 0.5 to obtain a binary segmentation. In Table 1, the Dice score, the Jaccard index, and the 95 percentile Hausdorff Distance (HD95) are presented. We achieve good results with respect to all those metrics.
For every image of the test set, we sample 5 different segmentation masks. This implicitly defines an ensemble by averaging over the 5 masks and thresholding it at 0.5. We report the results for this ensemble in the second line of Table 1. We see that already an ensemble of 5 increases the performance of our approach.
In the last column of Table 1, we count the cases where the model produces an empty segmentation mask. This results in a Dice of zero, and HD95 cannot be computed. If we disregard those cases, we report the HD95 score, and the average Dice score and Jaccard index are reported in square brackets in Table 1.
As baseline, we report the segmentation scores for the nnU-Net and SegNet. By default, nnU-Net is an ensemble of a 5-fold cross validation. We also implement Bayesian SegNet with Monte Carlo dropout as proposed in (Kendall et al., 2017). By sampling five times during inference, we can again make an ensemble of the generated segmentation masks. The scores for this ensemble are reported in the last line of Table 1.
The generation of one sample with our method takes 48 seconds, while the computation of the segmentation mask with SegNet takes 13 ms. To speed up the sampling process, we will consider sampling with the DDIM approach in future work.
For visualization of the uncertainty maps, we select three exemplary images b_1, b_2, and b_3 from the test set. More examples are presented in Appendix C. To generate detailed

Table 1: Segmentation scores of our method and nnU-Net on different metrics.

Method	**Dice**	**HD95**	**Jaccard**	**empty**
Ours (1 sampling run)	0.866 [0.892]	6.052	0.795 [0.819]	31
Ours (ensemble of 5 runs)	0.881 [0.909]	5.178	0.819 [0.845]	34
nnU-Net (ensemble of 5-fold cross-val.)	0.891 [0.905]	5.004	0.831 [0.845]	17
SegNet (1 run)	0.839 [0.867]	7.190	0.761 [0.786]	34
Bayesian SegNet (ensemble of 5 runs)	0.838 [0.841]	13.707	0.747 [0.749]	3

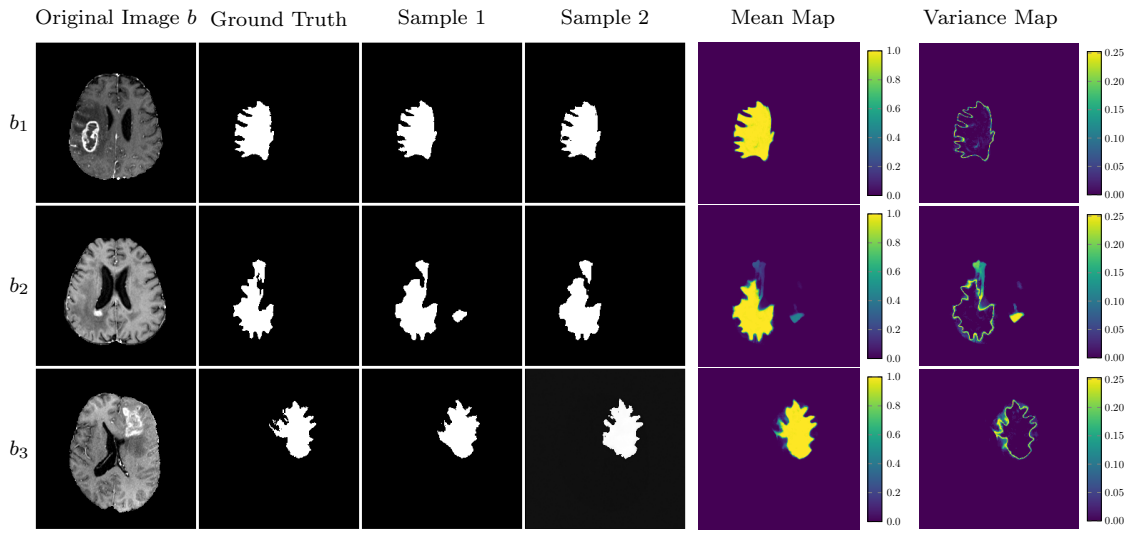


Figure 3: Examples of the produced mean and variance maps for 100 sampling runs.

uncertainty maps, we sample 100 segmentation masks for each of the images, and compute the pixel-wise variance. In Figure 3, we present one channel of the original brain MR image b, the ground truth segmentation, two different sampled segmentation masks, as well as the mean and variance map. We can clearly identify the areas where the model was uncertain. Moreover, by thresholding the mean map at 0.5, we can produce the ensembled segmentation mask. In Table 2, we report the segmentation scores and for this ensemble mask, as well as the average scores for the 100 samples. We see that the ensemble can boost the performance for the examples b_1, b_2 and b_3.

Table 2: Segmentation scores for the 100 samples of the examples presented in Figure 3.

	Average			**Ensemble**		
Example	**Dice**	**HD95**	**Jaccard**	**Dice**	**HD95**	**Jaccard**
b_1	0.969	2.360	0.939	0.981	1.000	0.962
b_2	0.869	18.503	0.769	0.885	18.468	0.783
b_3	0.932	5.227	0.872	0.952	4.474	0.907

In Figure 4, we plot the number of samples in the ensemble against the Dice score for the three examples b_1, b_2, and b_3. We can see that already an ensemble of five samples improves the performance, and then the curve flattens. In (Amit et al., 2021), a similar experiment was performed on a different data set. Independently from each other, we got the same findings. In Figure 5, we compare our variance maps against the ones of the Bayesian SegNet with Monte Carlo (MC) dropout for 100 samples, as well as the aleatoric uncertainty maps for SegNet, computed as proposed in (Kendall and Gal, 2017).

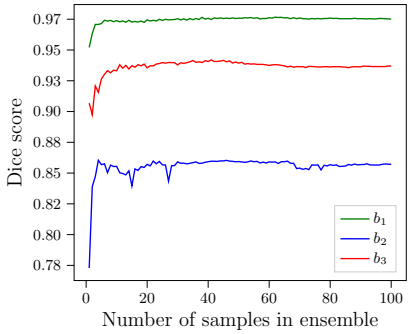


Figure 4: Performance of the ensemble with respect to the number of samples for the examples b_1, b_2, and b_3, presented in Figure 3.

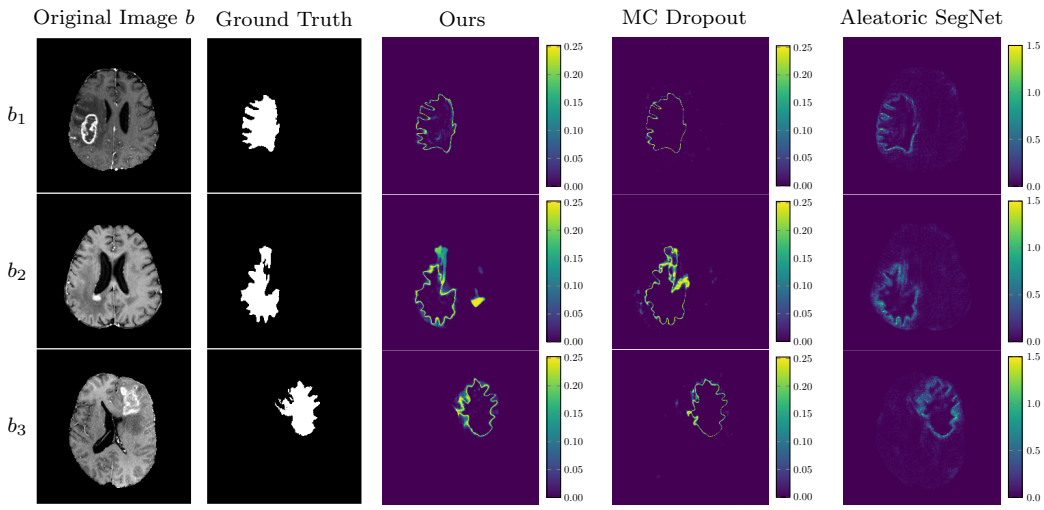


Figure 5: Comparison of the different uncertainty maps for the three examples.

5. Conclusion

We presented a novel approach for biomedical image segmentation based on DDPMs. Using the stochastic sampling process, our method allows implicit ensembling of different segmentation masks for the same input brain MR image, without having to train a new model. We could show that ensembling those segmentation masks increases the performance of the model with respect to different segmentation scores. Moreover, we can generate uncertainty maps by computing the variance of the different segmentation masks. This is of great interest in clinical applications, when we want to measure the uncertainty of the decision of the model. For future work, we plan to investigate the segmentation of the different tumor classes provided by the BRATS2020 challenge. Furthermore, we plan to use the DDIM scheme to speed up the sampling process.

Acknowledgments

This research was supported by the Novartis FreeNovation initiative and the Uniscientia Foundation (project # 147-2018).

References

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 2021.

Tomer Amit, Eliya Nachmani, Tal Shaharbany, and Lior Wolf. Segdiff: Image segmentation with diffusion probabilistic models. *arXiv preprint: arxiv:2112.00390*, 2021.

Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Scientific data*, 4(1):1–13, 2017.

Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. *arXiv preprint arXiv:1811.02629*, 2018.

Dmitry Baranchuk, Andrey Voynov, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. In *International Conference on Learning Representations*, 2022.

Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938*, 2021.

Terrance DeVries and Graham W Taylor. Leveraging uncertainty estimates for predicting segmentation quality. *arXiv preprint arXiv:1807.00502*, 2018.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34, 2021.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

Marija Habijan, Hrvoje Leventić, Irena Galić, and Danilo Babin. Whole heart segmentation from ct images using 3d u-net architecture. In *2019 International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 121–126. IEEE, 2019.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

Alain Jungo and Mauricio Reyes. Assessing reliability and challenges of uncertainty estimations for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 48–56. Springer, 2019.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 5580–5590. Curran Associates Inc., 2017.

Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. In *Proceedings of the British Machine Vision Conference*, pages 57.1–57.12. BMVA Press, September 2017.

Simon A. A. Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R Ledsam, Klaus Maier-Hein, SM Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. A probabilistic u-net for segmentation of ambiguous images. *Advances in neural information processing systems*, 31, 2018.

Simon A. A. Kohl, Bernardino Romera-Paredes, Klaus H Maier-Hein, Danilo Jimenez Rezende, SM Eslami, Pushmeet Kohli, Andrew Zisserman, and Olaf Ronneberger. A hierarchical probabilistic u-net for modeling multi-scale ambiguities. *arXiv preprint arXiv:1905.13077*, 2019.

Amish Kumar, Oduri Narayana Murthy, Palash Ghosal, Amritendu Mukherjee, Debashis Nandi, et al. A dense u-net architecture for multiple sclerosis lesion segmentation. In *TENCON 2019-2019 IEEE Region 10 Conference*, pages 662–667. IEEE, 2019.

Raghav Mehta, Angelos Filos, Yarin Gal, and Tal Arbel. Uncertainty evaluation metrics for brain tumour segmentation. In *Medical Imaging with Deep Learning*, 2020.

Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014.

John Mitros and Brian Mac Namee. On the validity of bayesian neural networks for uncertainty estimation. *arXiv preprint arXiv:1912.01530*, 2019.

Miguel Monteiro, Loïc Le Folgoc, Daniel Coelho de Castro, Nick Pawlowski, Bernardo Marques, Konstantinos Kamnitsas, Mark van der Wilk, and Ben Glocker. Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty. *Advances in Neural Information Processing Systems*, 33:12756–12767, 2020.

Tanya Nair, Doina Precup, Douglas L Arnold, and Tal Arbel. Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. *Medical image analysis*, 59:101557, 2020.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 8162–8171. PMLR, 2021.

Intisar Rizwan I Haque and Jeremiah Neubert. Deep learning approaches to biomedical image segmentation. *Informatics in Medicine Unlocked*, 18:100297, 2020.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

Abhinav Sagar. Uncertainty quantification using variational inference for biomedical image segmentation. *arXiv preprint arXiv:2008.07588*, 2020.

Chitwan Saharia, William Chan, Huiwen Chang, Chris A Lee, Jonathan Ho, Tim Salimans, David J Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. *arXiv preprint arXiv:2111.05826*, 2021.

Hiroshi Sasaki, Chris G Willcocks, and Toby P Breckon. Unit-ddpm: Unpaired image translation with denoising diffusion probabilistic models. *arXiv preprint arXiv:2104.05358*, 2021.

Matthew F Sharrock, W Andrew Mould, Hasan Ali, Meghan Hildreth, Issam A Awad, Daniel F Hanley, and John Muschelli. 3d deep neural network segmentation of intracerebral hemorrhage: Development and validation for clinical trials. *Neuroinformatics*, 19 (3):403–415, 2021.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338:34–45, 2019.

Zhitao Xiao, Bowen Liu, Lei Geng, Fang Zhang, and Yanbei Liu. Segmentation of lung nodules using improved 3d-unet neural network. *Symmetry*, 12(11):1787, 2020.

Appendix A. Exemplary Image of BRATS2020

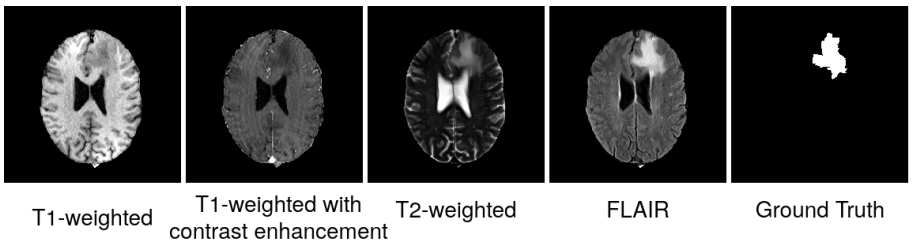


Figure 6: Exemplary image of the BRATS2020 dataset, with four different MR sequences and the ground truth segmentation.

Appendix B. Implementation Details

We provide implementation details of the comparing methods.

- SegNet: We train the SegNet as proposed in (Badrinarayanan et al., 2017), with a learning rate of 10^{-4} for the Adam optimizer and a batch size of 20. Training is performed with the binary cross-entropy loss and is stopped after 100 epochs.
- Bayesian SegNet: We adapt the SegNet architecture, and place the dropout layers with a dropout probability of $p = 0.5$ as proposed in (Kendall et al., 2017). The training schedule is kept the same as for SegNet.
- nnU-Net: We take over all hyperparameter settings as proposed in their official implementation, which can be found at `https://github.com/MIC-DKFZ/nnUNet`.
- Aleatoric Uncertainty Estimation: We keep the training settings for SegNet. The only change we need to make to the SegNet architecture is to double the number of output channels, such that we get both a prediction and a variance map. We follow the aleatoric loss implementation as proposed in (Jungo and Reyes, 2019), which can be found at `https://github.com/alainjungo/reliability-challenges-uncertainty`.

Appendix C. Further Examples

In Figure 7, we provide the mean and variance maps of three more exemplary images b_4, b_5, and b_6 of the test set.

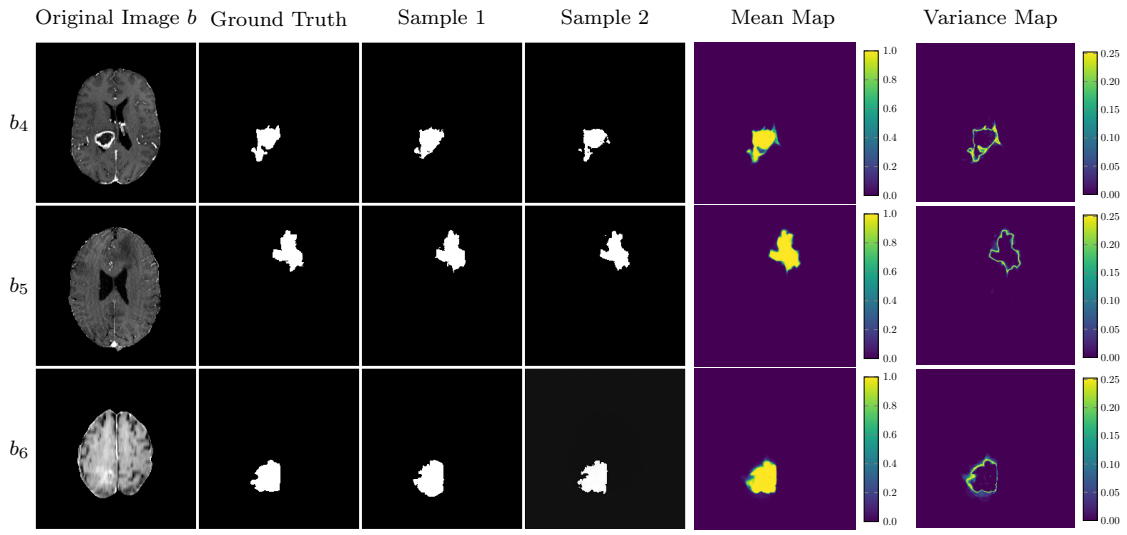


Figure 7: Additional examples of the produced mean and variance maps for 100 sampling runs.

Chapter 6

DeScarGAN: Disease-Specific Anomaly Detection with Weak Supervision

In this paper, we present a weakly supervised anomaly detection algorithm that is based on generative adversarial networks. We perform image-to-image translation between a set of images of patients and a set of healthy controls. By translating an image of a patient to a fake image of a healthy subject, the difference map highlights anomalous changes. We propose an architecture with weight sharing, skip connections and add an identity loss to ensure that only anomalous image regions are changed. This results in a detailed anomaly map. We focus on diseases where the anomaly exhibits a deformation of existing structures, which is different from detecting lesions. For the evaluation, we design a synthetic dataset that simulates such deformations and provides pixel-wise ground truth. We apply our method to the CheXpert dataset to detect pleural effusions in X-ray images of the lungs.

Publication. The following paper was presented at the *23d International Conference on Medical Image Computing and Computer Assisted Intervention* (MICCAI), October 2020, which was held virtually. It was published as part of the conference proceedings [198]. The code is publicly available[1].

[1] https://gitlab.com/cian.unibas.ch/DeScarGAN

DeScarGAN: Disease-Specific Anomaly Detection with Weak Supervision

Julia Wolleb, Robin Sandkühler, and Philippe C. Cattin

Department of Biomedical Engineering, University of Basel, Allschwil, Switzerland
julia.wolleb@unibas.ch

Abstract. Anomaly detection and localization in medical images is a challenging task, especially when the anomaly exhibits a change of existing structures, e.g., brain atrophy or changes in the pleural space due to pleural effusions. In this work, we present a weakly supervised and detail-preserving method that is able to detect structural changes of existing anatomical structures. In contrast to standard anomaly detection methods, our method extracts information about the disease characteristics from two groups: a group of patients affected by the same disease and a healthy control group. Together with identity-preserving mechanisms, this enables our method to extract highly disease-specific characteristics for a more detailed detection of structural changes. We designed a specific synthetic data set to evaluate and compare our method against state-of-the-art anomaly detection methods. Finally, we show the performance of our method on chest X-ray images. Our method called DeScarGAN outperforms other anomaly detection methods on the synthetic data set and by visual inspection on the chest X-ray image data set.

Keywords: Anomaly detection · Weak supervision · Disease-specific

1 Introduction

For medical applications, it is of great interest to find an automated way to show visual manifestations of a disease. In the past, artificial neural networks have shown a great performance in the task of image segmentation. As the manual generation of pixel-wise annotations is time consuming and requires expert knowledge, the training data is limited in number or even unavailable. Furthermore, the manually generated labels are affected by human bias. Using only image-level class labels for training of the networks overcomes those issues. In this paper, we propose a new disease-specific and weakly supervised method for anomaly detection and localization. The task we aim to solve is to highlight the pathological changes in an image of a diseased subject, as well as the classification into diseased and healthy subjects. This can improve diagnosis, lead the attention to relevant parts of the anatomy and provide a starting point for further studies.

Classical anomaly detection algorithms are trained only on healthy subjects and detect abnormal parts of images as outliers. Variational Autoencoders

(VAEs) can be used to detect lesions in the brain [4, 25]. Beside VAEs, Generative Adversarial Networks (GANs) [8] are used for anomaly detection in medical images [7]. VAGAN [3] proposes the generation of an additive map to make an image of a diseased subject appear healthy. PathoGAN [2] provides a weakly supervised segmentation algorithm for brain tumors based on image-to-image translation. StarGAN [5] follows a similar idea as CycleGAN [24] and simplifies the architecture to only one generator and one discriminator. This idea can be used for anomaly detection by taking the difference between original and translated images. Fixed-Point GAN (FP-GAN) [20] improves StarGAN by preserving features that should not be changed during translation, outperforming f-Anogan [18] and others [1] in brain lesion detection. The problem of combining GANs with a classification network is tackled by semi-supervised GANs [15, 17]. Class activation maps [19, 23] visualize the features of the input image that lead to the classification score, but limitations in the resolution lead to blurry maps. Another approach is the generation of saliency maps [13, 21] by computing the gradient of the classification score with regard to the input image.
We are interested in cases where the anomaly occurs in the form of deformations of existing structures, e.g., atrophy, rather than in lesions. Both VAGAN and FP-GAN are designed to only generate an additive map rather than a complete new image. We claim that this restriction to additive maps may hinder the methods from showing deformations. What is more, VAGAN is not designed to perform classification and assumes that the class label for each input image is provided in advance. VAEs are only trained on the healthy control group and may not be able to point out the characteristics of a specific disease, due to natural variations in the data. Our method for detection of structural changes in anatomical regions, further called DeScarGAN, is designed to address these issues.

Our method performs image-to-image translation between a set of healthy and a set of diseased subjects in order to find the visual manifestations that make the distributions of the two datasets differ from each other. We introduce a novel disease-specific architecture with skip connections, a splitting of the networks into weight-sharing subnetworks and an identity loss as identity-preserving mechanisms. This ensures that the difference between the generated healthy and the real input image is accurate enough to highlight the regions of interest, resulting in more detailed maps of the characteristics of the disease than previous methods.
We point out that compared to classical anomaly detection, we train only on one specific disease and extract information about its characteristics. With this approach, changes of already existing structures can be detected in a detailed manner, which is different from the presence or absence of lesions.
We evaluate our method on a synthetic dataset designed for this task. Furthermore, we apply it on the Chexpert dataset [12] of X-ray images of lungs in order to detect pleural effusions. Our method outperforms state-of-the-art anomaly detection algorithms in showing deformations of already existing structures. Furthermore, it provides better classification results than standard clas-

sification algorithms. With the addition of visually highlighting the regions of interest, the attention is led to the relevant parts of the image, making a step towards *interpretable machine learning*. The code is publicly available at `https://github.com/JuliaWolleb/DeScarGAN`.

2 Method

Let $\mathcal{F} = \{x \mid x : \mathbb{R}^2 \to \mathbb{R}\}$ be a set of medical images from the same imaging modality showing the same anatomical structures, with $\mathcal{P} \subset \mathcal{F}$ the set of images of patients affected by a specific disease and $\mathcal{H} \subset \mathcal{F}$ the set of images of a healthy control group. The aim of our method is, given a new image of unknown class, to detect regions in the image that show the same characteristics as the images in $\mathcal{P}$ and to assign a class label.
Let p be the class of the images in $\mathcal{P}$ and h the class of the images in $\mathcal{H}$, with $c, \bar{c} \in \{h, p\}$ and $c \neq \bar{c}$. The main idea is to translate a real image r_c of either class c to an artificial image $a_{\bar{c}}$ of class $\bar{c}$. The pathological region is then defined as the difference $d := a_h - r_c$ between the artificial healthy image a_h and the real input image r_c of class c. Thus we perform image-to-image translation between the unpaired sets $\mathcal{P}$ and $\mathcal{H}$. A diagram showing the workflow of our method is given in Figure 1. Given any image r_c, the generator both generates an artificial image a_c of the same class c and an artificial image $a_{\bar{c}}$ of class $\bar{c}$. To ensure that r_c and a_h only differ in the pathological region, we add the identity loss $\mathcal{L}_{id}$ and the reconstruction loss $\mathcal{L}_{rec}$ for cycle consistency.

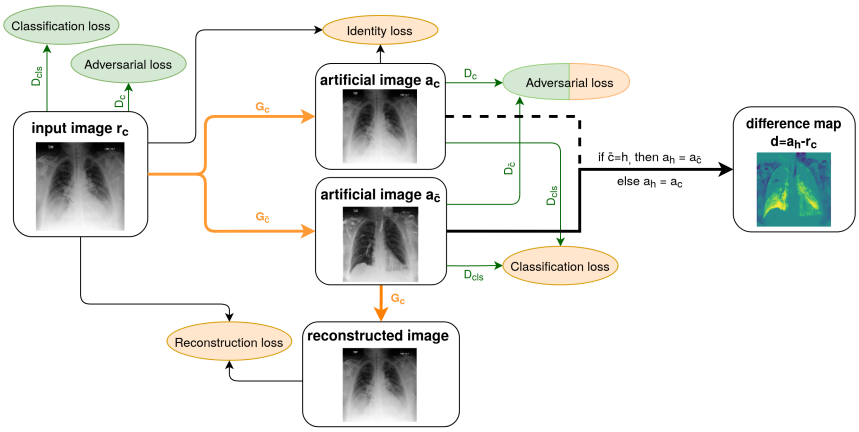


Fig. 1. Workflow of our method. The components of the loss functions for the discriminator are shown in green, the ones for the generator in orange.

The generator consists of two branches, its architecture is shown in Figure 2. We refer to the generator $G_p : \mathcal{F} \to \mathcal{P}$ for generating images of class p and generator $G_h : \mathcal{F} \to \mathcal{H}$ for generating images of class h.

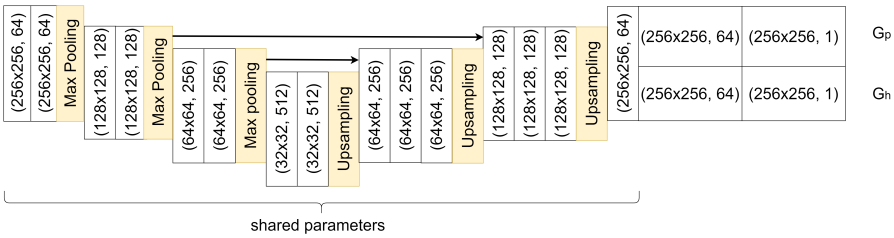


Fig. 2. The architecture of the generator network. Every box stands for a convolutional layer with the stated output size (image width × image height, feature channels) and kernelsize 3, followed by a batch normalization layer and a ReLU activation function.

The skip connections of the generator ensure that the artificial image maintains the detailed structures of the input image. This is a way to alter only the necessary features, thus making the difference map d more accurate. The skip connection in the uppermost layer turned out to be too restrictive to perform the translation to another class. By omitting this skip connection, we enable the generator to perform structural changes.

The discriminator network has the task to both classify images into healthy and diseased subjects and to distinguish between real and artificial images. Therefore, it consists of three subnets that share parameters, as shown in Figure 3. $D_p : \mathcal{P} \to \mathbb{R}$ distinguishes between real and artificial images of class p, $D_h : \mathcal{H} \to \mathbb{R}$ does the same for class h and $D_{cls} : \mathcal{F} \to \mathbb{R}$ is the network for classification, following the structure of a VGG net [22]. The branching of the generator and discriminator gives a higher range of flexibility compared to StarGAN, which turned out to be beneficial for image-to-image translation.

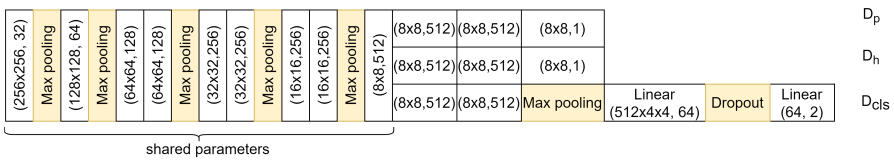


Fig. 3. The architecture of the discriminator with three subnets D_p, D_h and D_{cls} that share parameters. Every box stands for a convolutional layer with kernelsize 3 and with the stated output size, followed by a ReLU activation function.

With the notation from above, D_c can be D_p or D_h interchangeably, and $D_{\bar{c}}$ denotes the discriminator for the contrary class. The same applies for the generator G.

2.1 Loss functions

Adversarial Loss The generator aims to generate images that the discriminator cannot distinguish from real images. Following the idea of Wasserstein

GANs [9], we add a gradient penalty loss and define the adversarial loss for the discriminator as

$$\mathcal{L}_{adv,d} = -\mathbb{E}_{r_c,c}[(D_c(r_c))] + \mathbb{E}_{r_c,\bar{c}}[D_{\bar{c}}(G_{\bar{c}}(r_c)] + \lambda_{gp}\mathbb{E}_{\hat{x},c}[(\parallel \nabla_{\hat{x}} D_c(\hat{x}_c) \parallel_2 -1)^2], \tag{1}$$

where $\hat{x}_c$ is given by $\hat{x}_c = tr_c + (1-t)a_c$ with $t \sim U([0,1])$. The adversarial loss for the generator is defined as

$$\mathcal{L}_{adv,g} = -\mathbb{E}_{r_c,\bar{c}}[D_{\bar{c}}(G_{\bar{c}}(r_c)]. \tag{2}$$

Identity Loss Considering an input image r_c, we aim for identity between r_c and $G_c(r_c)$. Therefore, the identity loss for the generator is defined as

$$\mathcal{L}_{id} = \mathbb{E}_{r_c,c}[\parallel r_c - G_c(r_c) \parallel_2]. \tag{3}$$

Classification Loss The classification subnet D_{cls} of the discriminator has to correctly classify r_c to belong to class c. The objective function for the discriminator is described as

$$\mathcal{L}_{cls,d} = \mathbb{E}_{r_c,c}[-\log D^c_{cls}(r_c)], \tag{4}$$

where the term $D^c_{cls}(r_c)$ describes the computed probability score that r_c belongs to class c. The generator aims for classification of an artificial image $a_{\bar{c}} = G_{\bar{c}}(r_c)$ to belong to class $\bar{c}$. Therefore, the classification loss for the generator is defined as

$$\mathcal{L}_{cls,g} = \mathbb{E}_{r_c,\bar{c}}[-\log D^{\bar{c}}_{cls}(G_{\bar{c}}(r_c))]. \tag{5}$$

Reconstruction Loss When an input image r_c of class c is translated into an image $a_{\bar{c}} = G_{\bar{c}}(r_c)$ of class $\bar{c}$, we aim for cycle consistency when translating $a_{\bar{c}}$ back to class c. This is achieved by adding a reconstruction loss term for the generator, given by

$$\mathcal{L}_{rec} = \mathbb{E}_{r_c,c}[\parallel r_c - G_c(G_{\bar{c}}(r_c)) \parallel_2]. \tag{6}$$

Total Loss Objective The overall loss function for the generator is defined as

$$\mathcal{L}_g = \lambda_{adv,g}\mathcal{L}_{adv,g} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{id}\mathcal{L}_{id} + \lambda_{cls,g}\mathcal{L}_{cls,g}, \tag{7}$$

and for the discriminator as

$$\mathcal{L}_d = \lambda_{adv,d}\mathcal{L}_{adv,d} + \lambda_{cls,d}\mathcal{L}_{cls,d}. \tag{8}$$

3 Synthetic Dataset

The purpose of weakly supervised algorithms is to overcome the need for pixel-wise labels and the human bias within these labels. In order not to be affected by this human bias, we designed a synthetic data set for the evaluation of our method. Two ellipses e_1 and e_2 are present in the image, one larger than the other and both with variable contour thickness, origin and orientation. The background is structured in concentric waves with two variable origins and variable wave length; this provides a higher level of complexity. Images of the healthy group $\mathcal{H}$ keep this structure. If the image is deformed such that the smaller ellipse e_1 shrinks to an even smaller ellipse, the background is also deformed. Images with this characteristics belong to the diseased group $\mathcal{P}$. Implementation details are provided in the supplementary material.
In Figure 4, exemplary images of the two sets $\mathcal{H}$ and $\mathcal{P}$ are shown. The pixel-wise ground truth (GT) is known by definition. We generate a training set of 2000 images of each class, and a validation and a test set with 200 images of each class.

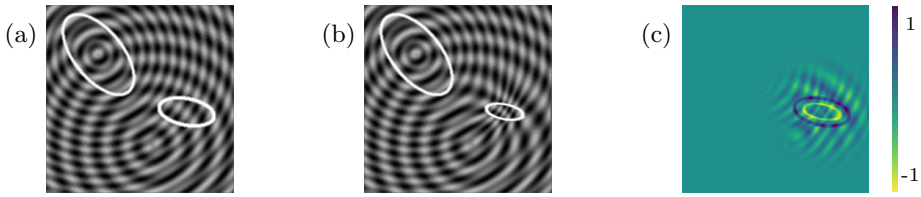


Fig. 4. Images (a) and (b) show exemplary images of the sets $\mathcal{H}$ and $\mathcal{P}$ respectively. Image (c) corresponds to the ground truth given by the difference (a) - (b).

4 Results and Discussion

We compare our method against StarGAN, FP-GAN, the VAE proposed in [4] and VAGAN. To train our model, we use the Adam optimizer [14] with $\beta_1 = 0.5$, $\beta_2 = 0.999$, and a learning rate of 10^{-4}. For every update of the parameters of the generator, we update the discriminator 5 times. We manually choose the hyperparameters $\lambda_{adv,d} = 20$, $\lambda_{gp} = 10$, $\lambda_{id} = \lambda_{rec} = 50$, $\lambda_{adv,g} = \lambda_{cls,g} = 1$, and $\lambda_{cls,d} = 5$. The number of trained parameters is 8528262 for the generator and 18170180 for the discriminator.

4.1 Synthetic Dataset

As a measure for the pixel-wise error for the anomaly detection task, we choose the Dice score, AUROC_{pix} [10] for pixel-wise classification, the Mean Square Error (MSE) and the Structural Similarity Index (SSIM) between d and GT,

and finally the MSE between an input image $r_h \in \mathcal{H}$ and the corresponding artificial image a_h. For the calculation of the Dice score and the AUROC_{pix}, we perform a thresholding based on the average Otsu [16] threshold value on the GT images. The results are shown in Table 1. All methods classify the images almost perfectly on the test set, so we omit those results. VAGAN is not designed to take an image $r_h \in \mathcal{H}$ as input, but we still report the result for completeness.

Table 1. Results on the synthetic dataset.

	Dice	AUROC_{pix}	$\text{MSE}(d, GT)$ (var)	SSIM	$\text{MSE}(r_h, a_h)$ (var)
StarGAN	0.710	0.962	0.0229 (0.128)	0.888	0.0025 (0.002)
FP-GAN	0.766	0.975	0.0160 (0.004)	0.917	0.0027 (0.003)
VAGAN	0.442	0.954	0.1321 (0.132)	0.869	0.0036 (0.002)
VAE	0.288	0.809	0.0734 (0.071)	0.668	0.0316 (0.031)
DeScarGAN	**0.853**	**0.988**	**0.0086(0.002)**	**0.954**	**0.0018 (0.001)**

In Figure 5, exemplary real images $r_p \in \mathcal{P}$ of the synthetic dataset with the corresponding artificial images $a_h \in \mathcal{H}$ of the different methods are shown. Our method provides the most accurate difference map d. The results of FP-GAN are good as well, but the method fails to generate a proper unshrunken ellipse e_1. VAE and VAGAN fail to generate an accurate image of class h, resulting in a difference map not close to the ground truth. For visualization, we omit the StarGAN method since it is outperformed by its extension FP-GAN.

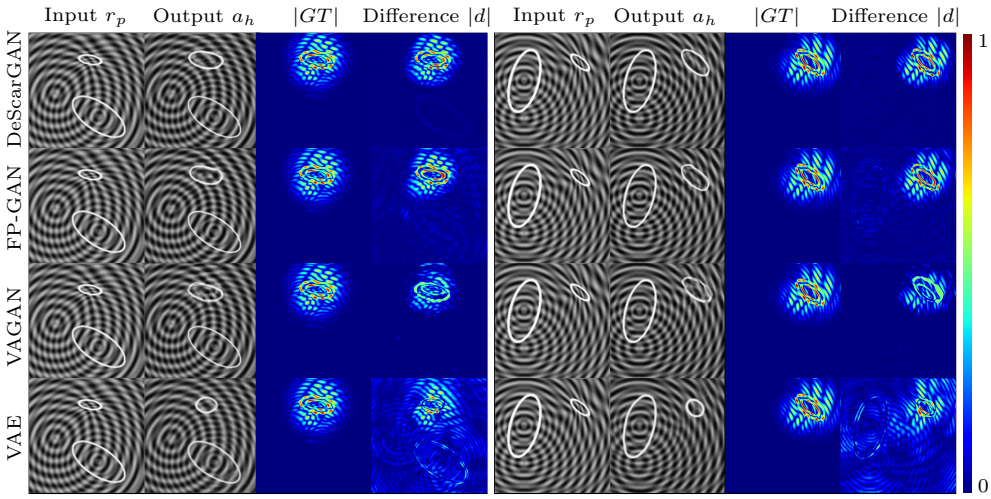


Fig. 5. Visualization of the results of our DeScarGAN, FP-GAN, VAGAN and VAE for two samples of the synthetic dataset.

4.2 Chexpert Dataset

For the Chexpert dataset introduced in [12], we used a training set of 14179 images of healthy subjects and 16776 images of subjects that suffer from pleural effusions. The test and validation set each consist of 200 images for each class.

Table 2. Classification results and MSE(r_h, a_h) on the Chexpert dataset.

	Accuracy$_{cls}$	Kappa score	AUROC$_{image}$	MSE(r_h, a_h) (var)
StarGAN	0.853	0.705	0.923	0.0534 (0.095)
FP-GAN	0.875	0.750	0.939	0.0060 (0.007)
VAGAN	×	×	×	0.0638 (0.065)
VAE	×	×	×	0.0231 (0.030)
Densenet169	0.893	0.785	0.951	×
D_{cls}	0.890	0.780	0.949	×
DeScarGAN	**0.898**	**0.795**	**0.953**	**0.0035 (0.003)**

For classification, we compare DeScarGAN against the classification results of StarGAN, FP-GAN, Densenet169 [11] and the classifier D_{cls} without the GAN mechanism. The result for the image-level classification is measured in classification accuracy, the Cohen's kappa score [6] and the AUROC score. Further, we measure the MSE between real images $r_h \in \mathcal{H}$ and artificial images $a_h \in \mathcal{H}$. The scores are summarized in Table 2.

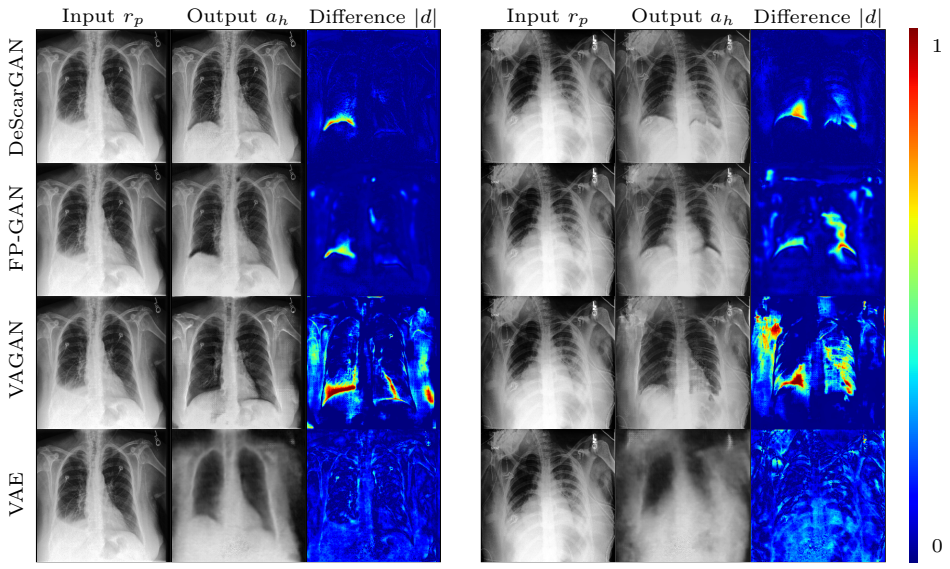


Fig. 6. Comparison of our DeScarGAN against FP-GAN, VAGAN and VAE for two samples of the Chexpert dataset.

DeScarGAN achieves better classification results than the pure classification networks D_{cls} and Densenet169, indicating that the GAN mechanism supports the classification network. The results of the different methods are visualized in Figure 6. We observe that the VAE fails to detect pleural effusions. Although FP-GAN detects similar regions as our method, the generated maps appear blurry and mark regions outside the thorax. The additive map of VAGAN also outlines parts of the arms and upper chest as abnormal. Our method generates the most detailed difference map, not highlighting any regions outside the pleural space.

5 Conclusion

We proposed DeScarGAN, a method to generate disease-specific, detailed maps that show pathological changes of existing anatomical structures. The novelty of our method is the introduction of a new architecture with skip connections, a splitting of the networks into weight-sharing subnetworks and an identity loss as identity-preserving mechanisms. This setup enables the detection of deformations of existing anatomical structures, e.g., atrophy or changes in the pleural space due to pleural effusions.
When comparing our DeScarGAN against state-of-the-art anomaly detection algorithms, we outperform FP-GAN, VAE, StarGAN and VAGAN on a synthetic dataset. Although FP-GAN provides good results by generating additive maps, DeScarGAN generates a complete new image and provides more precise maps that reliably outline the regions of pathological changes.
When applying our model on the Chexpert lung X-ray dataset with pleural effusions, our classification scores are better than state-of-the-art classification networks. The generated maps detect anomalies in a detailed manner and lead the attention to the relevant parts of the anatomy. This approach has the potential to bridge the gap between the knowledge about the presence of a disease and setting the focus of a longitudinal study observing the region of interest.

Acknowledgement. This work was supported by Novartis FreeNovation.

References

1. Alex, V., Safwan, K.P.M., Chennamsetty, S.S., Krishnamurthi, G.: Generative adversarial networks for brain lesion detection. In: Medical Imaging 2017: Image Processing. vol. 10133, pp. 113 – 121. International Society for Optics and Photonics, SPIE (2017)
2. Andermatt, S., Horváth, A., Pezold, S., Cattin, P.: Pathology segmentation using distributional differences to images of healthy origin. In: International MICCAI Brainlesion Workshop. pp. 228–238. Springer (2018)
3. Baumgartner, C.F., Koch, L.M., Can Tezcan, K., Xi Ang, J., Konukoglu, E.: Visual feature attribution using wasserstein GANs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8309–8319 (2018)

4. Chen, X., Konukoglu, E.: Unsupervised detection of lesions in brain MRI using constrained adversarial auto-encoders. arXiv preprint arXiv:1806.04972 (2018)
5. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8789–8797 (2018)
6. Cohen, J.: A coefficient of agreement for nominal scales. Educational and Psychological Measurement **20**(1), 37–46 (1960)
7. Di Mattia, F., Galeone, P., De Simoni, M., Ghelfi, E.: A survey on GANs for anomaly detection. arXiv preprint arXiv:1906.11632 (2019)
8. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems 27, pp. 2672–2680 (2014)
9. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein GANs. In: Advances in Neural Information Processing Systems 30. pp. 5767–5777 (2017)
10. Hanley, J.A., McNeil, B.J.: The meaning and use of the area under a receiver operating characteristic (roc) curve. Radiology **143**(1), 29–36 (1982)
11. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition. pp. 2261–2269 (2017)
12. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R.L., Shpanskaya, K.S., Seekins, J., Mong, D.A., Halabi, S.S., Sandberg, J.K., Jones, R., Larson, D.B., Langlotz, C.P., Patel, B.N., Lungren, M.P., Ng, A.Y.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 590–597 (2019)
13. Karargyros, A., Syeda-Mahmood, T.: Saliency u-net: A regional saliency map-driven hybrid deep learning network for anomaly segmentation. In: Medical Imaging 2018: Computer-Aided Diagnosis. vol. 10575, pp. 413 – 418. International Society for Optics and Photonics, SPIE (2018)
14. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
15. Odena, A.: Semi-supervised learning with generative adversarial networks. arXiv preprint arXiv:1606.01583 (2016)
16. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man, and Cybernetics **9**(1), 62–66 (1979)
17. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training GANs. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. pp. 2234–2242 (2016)
18. Schlegl, T., Seeböck, P., Waldstein, S.M., Langs, G., Schmidt-Erfurth, U.: f-anogan: Fast unsupervised anomaly detection with generative adversarial networks. Medical Image Analysis **54**, 30–44 (2019)
19. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International Conference on Computer Vision. pp. 618–626 (2017)
20. Siddiquee, M.M.R., Zhou, Z., Tajbakhsh, N., Feng, R., Gotway, M.B., Bengio, Y., Liang, J.: Learning fixed points in generative adversarial networks: From image-to-image translation to disease detection and localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 191–200 (2019)

21. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)
22. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (2015)
23. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition. pp. 2921–2929 (2016)
24. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. 2017 IEEE International Conference on Computer Vision pp. 2242–2251 (2017)
25. Zimmerer, D., Isensee, F., Petersen, J., Kohl, S., Maier-Hein, K.: Unsupervised anomaly localization using variational auto-encoders. Medical Image Computing and Computer Assisted Intervention pp. 289–297 (2019)

Supplementary Material to DeScarGAN

Julia Wolleb, Robin Sandkühler, and Philippe C. Cattin

1 Generation of the Synthetic Dataset

Table 1. Equations used for the generation of the synthetic dataset. The images are of size 256×256, and the image coordinates run from 0 to 255. The ellipses have a rotation angle ϕ and a contour thickness g. If the two ellipses intersect, the image is excluded from the dataset. The images of the healthy group $\mathcal{H}$ show the two ellipses e_1 and e_2, the background is given by the circles c_1 and c_2. The images of the diseased group $\mathcal{P}$ are generated by deforming a healthy image with the given deformation field.

name	formula	parameters
ellipse e_1	$(a\cos(\theta) + m_1, b\sin(\theta) + m_2) = 1$	$m_1, m_2 \sim U([50, 200])$, $\forall\theta \in [0, 2\pi)$, $a \sim \mathcal{N}(40, 1)$, $b \sim \mathcal{N}(20, 1)$, $\phi \sim U([0, 2\pi])$, $g \sim \mathcal{N}(5, 0.7)$
ellipse e_2	$(a\cos(\theta) + m_1, b\sin(\theta) + m_2) = 1$	$m_1, m_2 \sim U([50, 200])$, $\forall\theta \in [0, 2\pi)$, $a \sim \mathcal{N}(70, 1)$, $b \sim \mathcal{N}(35, 1)$, $\phi \sim U([0, 2\pi])$, $g \sim \mathcal{N}(5, 0.7)$
circle c_1	$(t\cos\theta + m1, t\sin\theta + m_2) = \sin(tf)h$	$\forall t \in [0, 255]$, $\forall\theta \in [0, 2\pi)$, $h \sim U([0.2, 0.4])$, $f \sim U([0.2, 0.35])$, $m_1, m_2 \sim U([50, 200])$
circle c_2	$(t\cos\theta + m1, t\sin\theta + m_2) = \sin(tf)h$	$\forall t \in [0, 255]$, $\forall\theta \in [0, 2\pi)$, $h \sim U([0.2, 0.4])$, $f \sim U([0.35, 0.5])$, m_1 and m_2 of e_2
deformation field	$\nabla\left(\frac{1}{(2\pi * 0.19^2)} e^{-\left(\frac{(x-m_1)^2}{(2*0.19^2)} + \frac{(y-m_2)^2}{(2*0.19^2)}\right)}\right)$	m_1 and m_2 of e_1

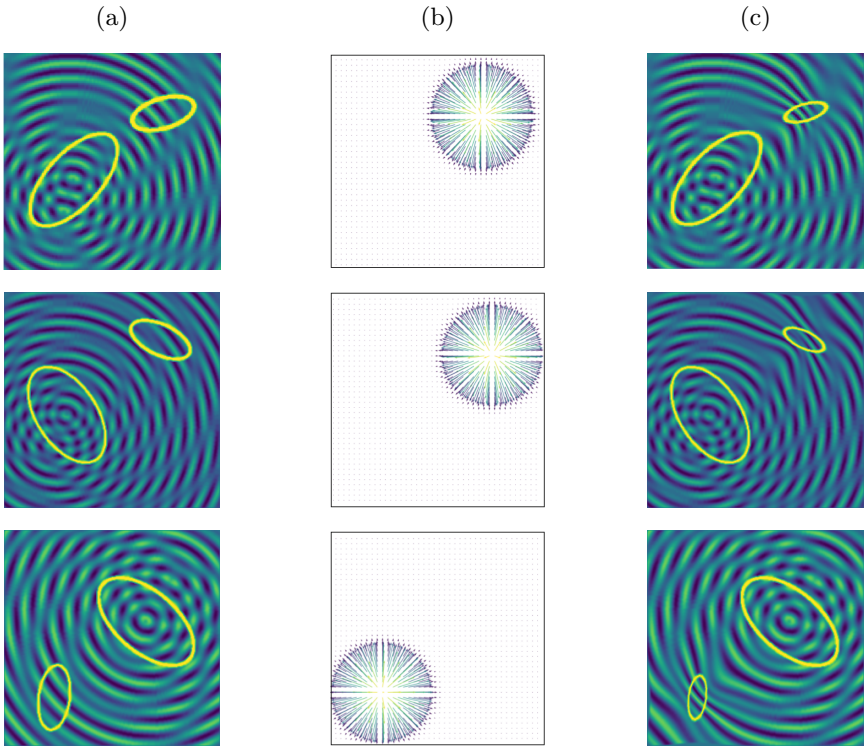


Fig. 1. Image (a) shows an exemplary image of the set $\mathcal{H}$. Image (b) shows the deformation field used to transform image (a) into image (c) of the diseased group $\mathcal{P}$. For the dataset, either image (a) or image (c) is chosen, such that the sets $\mathcal{H}$ and $\mathcal{P}$ remain unpaired.

Chapter 7

Diffusion Models for Medical Anomaly Detection

Since the training of GANs is cumbersome and unstable, we replace the GAN in the weakly supervised anomaly detection approach proposed in Chapter 6 with a denoising diffusion model. Given two datasets, one containing images of healthy controls and one containing images of patients, the goal is to translate images of patients into synthetic images of healthy controls. The diffusion model is trained to generate realistic-looking images. A separate binary classification model is trained to distinguish between healthy and diseased subjects. By using the iterative deterministic noising and denoising scheme of denoising diffusion implicit models, an input image can be encoded in noise and translated to the desired output class using gradient guidance. Thereby, only the anomalous image regions are changed, rendering a very detailed difference map between an input image of a patient and the output image of a healthy subject.
We evaluate our approach on the CheXpert dataset for detecting pleural effusions and on the BRATS2020 dataset for brain tumor detection.

Publication. The following paper is accepted at the *25th International Conference on Medical Image Computing and Computer Assisted Intervention* (MICCAI), September 2022, Singapore. The manuscript was published as part of the conference proceedings [194]. The code for this framework is open-source[1].

[1] https://gitlab.com/cian.unibas.ch/diffusion-anomaly

Diffusion Models for Medical Anomaly Detection

Julia Wolleb, Florentin Bieder, Robin Sandkühler, Philippe C. Cattin

Department of Biomedical Engineering, University of Basel, Allschwil, Switzerland
julia.wolleb@unibas.ch

Abstract. In medical applications, weakly supervised anomaly detection methods are of great interest, as only image-level annotations are required for training. Current anomaly detection methods mainly rely on generative adversarial networks or autoencoder models. Those models are often complicated to train or have difficulties to preserve fine details in the image. We present a novel weakly supervised anomaly detection method based on denoising diffusion implicit models. We combine the deterministic iterative noising and denoising scheme with classifier guidance for image-to-image translation between diseased and healthy subjects. Our method generates very detailed anomaly maps without the need for a complex training procedure. We evaluate our method on the BRATS2020 dataset for brain tumor detection and the CheXpert dataset for detecting pleural effusions.

Keywords: Anomaly detection · Diffusion models · Weak supervision.

1 Introduction

In medical image analysis, pixel-wise annotated ground truth is hard to obtain, often unavailable and contains a bias to the human annotators. Weakly supervised anomaly detection has gained a lot of interest in research as an essential tool to overcome the aforementioned issues. Compared to fully supervised methods, weakly supervised models rely only on image-level labels for training. In this paper, we present a novel pixel-wise anomaly detection approach based on Denoising Diffusion Implicit Models (DDIMs) [25]. Figure 1 shows an overview

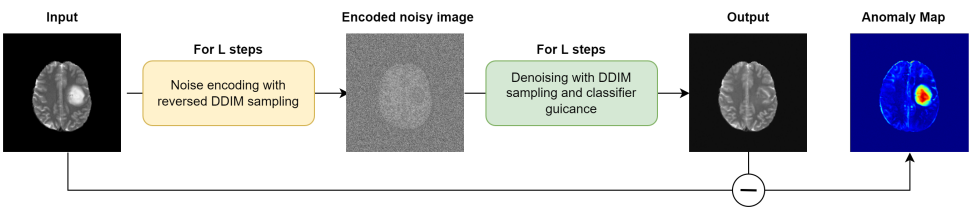


Fig. 1. Proposed sampling scheme for image-to-image translation between a diseased input image and a healthy output image. The anomaly map is defined as the difference between the two.

of the proposed method. We assume two unpaired sets of images for the training, the first containing images of healthy subjects and the second images of subjects affected by a disease. Only the image and the corresponding image-level label (healthy, diseased) are provided during training.
Our method consists of two main parts. In the first part, we train a Denoising Diffusion Probabilistic Models (DDPM)[10] and a binary classifier on a dataset of healthy and diseased subjects. In the second part, we create the actual anomaly map of an unseen image. For this, we first encode the anatomical information of an image with the reversed sampling scheme of DDIMs. This is an iterative noising process. Then, in the denoising process, we use the deterministic sampling scheme proposed in DDIM with classifier guidance to generate an image of a healthy subject. The final pixel-wise anomaly map is the difference between the original and the synthetic image. With this encoding and denoising procedure, our method can preserve many details of the input image that are not affected by the disease while re-painting the diseased part with realistic looking tissue. We apply our algorithm on two different medical datasets, i.e., the BRATS2020 brain tumor challenge [16, 2, 3], and the CheXpert dataset [11], and compare our method against standard anomaly detection methods. The source code and implementation details are available at `https://gitlab.com/cian.unibas.ch/diffusion-anomaly`.

Related Work In classical anomaly detection, autoencoders [29, 13] are trained on data of healthy subjects. Any deviations from the learned distribution then lead to a high anomaly score. This idea has been applied for unsupervised anomaly detection in medical images [30, 6, 14], where the difference between the healthy reconstruction and the anomalous input image highlight pixels that are perceived as anomalous. Other approaches focus on Generative Adversarial Networks (GANs) [9] for image-to-image translation [24, 5, 27].
However, training of GANs is challenging and requires a lot of hyperparameter tuning. Furthermore, additional loss terms and changes to the architecture are required to ensure cycle-consistent results. In [19, 1], the gradient of a classifier is used to obtain anomaly maps. Recently, transformer networks [21] were also successfully applied on brain anomaly detection [20]. Non-synthesis based methods such as density estimation, feature modeling or self-supervised classification also provide state-of-the-art techniques for anomaly detection [28]. In [15], a new thresholding method is proposed for anomaly segmentation on the BRATS dataset.
Lately, DDPMs were in focus for there ability to beat GANs on image synthesis [8]. In the flow of this success, they were also applied on image-to-image translation [23, 7], segmentation [4], reconstruction [22] and registration[12]. As shown in [25], DDIMs are closely related to score-based generative models [26], which can be used for interpolation between images. However, there is no diffusion model for anomaly detection so far to the best of our knowledge.

2 Method

A typical example for image-to-image translation in medicine is the transformation of an image of a patient to an image without any pathologies. For anomaly detection it is crucial that only pathological regions are changed, and the rest of the image is preserved. Then, the difference between the original and the translated image defines the anomaly map. Our detail-preserving image-to-image translation is based on diffusion models. We follow the formulation of DDPMs given in [10, 17]. In Algorithm 1, we present the workflow of our approach.
The general idea of diffusion models is that for an input image x, we generate a series of noisy images $\{x_0, x_1, ..., x_T\}$ by adding small amounts of noise for many timesteps T. The noise level t of an image x_t is steadily increased from 0 to T. A U-Net ϵ_θ is trained to predict x_{t-1} from x_t according to (5), for any step $t \in \{1, ..., T\}$. During training, we know the ground truth for x_{t-1}, and the model is trained with an MSE loss. During evaluation, we start from $x_T \sim \mathcal{N}(0, \mathbf{I})$ and predict x_{t-1} for $t \in \{T, ..., 1\}$. With this iterative denoising process, we can generate a a fake image x_0. The forward noising process q with variances $\beta_1, ..., \beta_T$ is defined by

$$q(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}). \tag{1}$$

This recursion can be written explicitly as

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \quad \text{with } \epsilon \sim \mathcal{N}(0, \mathbf{I}). \tag{2}$$

with $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$. The denoising process p_θ is learned by optimizing the model parameters θ and is given by

$$p_\theta(x_{t-1}|x_t) := \mathcal{N}\big(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)\big). \tag{3}$$

The output of the U-Net is denoted as ϵ_θ, and the MSE loss used for training is

$$\mathcal{L} := ||\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)||_2^2, \quad \text{with } \epsilon \sim \mathcal{N}(0, \mathbf{I}). \tag{4}$$

As shown in [25], we use the DDPM formulation to predict x_{t-1} from x_t with

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\left(\frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t, t)}{\sqrt{\bar{\alpha}_t}}\right) + \sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}\epsilon_\theta(x_t, t) + \sigma_t\epsilon, \tag{5}$$

with $\sigma_t = \sqrt{(1-\bar{\alpha}_{t-1})/(1-\bar{\alpha}_t)}\sqrt{1-\bar{\alpha}_t/\bar{\alpha}_{t-1}}$. DDPMs have a stochastic element ϵ in each sampling step (5). In DDIMs however, we set $\sigma_t = 0$, which results in a deterministic sampling process. As derived in [25], (5) can be viewed as the Euler method to solve an ordinary differential equation (ODE). Consequently, we can reverse the generation process by using the reversed ODE. Using enough discretization steps, we can encode x_{t+1} given x_t with

$$x_{t+1} = x_t + \sqrt{\bar{\alpha}_{t+1}}\left[\left(\sqrt{\frac{1}{\bar{\alpha}_t}} - \sqrt{\frac{1}{\bar{\alpha}_{t+1}}}\right)x_t + \left(\sqrt{\frac{1}{\bar{\alpha}_{t+1}} - 1} - \sqrt{\frac{1}{\bar{\alpha}_t} - 1}\right)\epsilon_\theta(x_t, t)\right]. \tag{6}$$

By applying (6) for $t \in \{0, ..., T-1\}$, we can encode an image x_0 in a noisy image x_T. Then, we recover the identical x_0 from x_T by using (5) with $\sigma_t = 0$ for $t \in \{T, ..., 1\}$.
For anomaly detection, we train a DDPM on a dataset containing images of healthy and diseased subjects. For evaluation, we define a noise level $L \in \{1, ..., T\}$ and a gradient scale s. Given an input image x, we encode it to a noisy image x_L using (6) for $t \in \{0, ..., L-1\}$. With this iterative noising process, we can induce anatomical information of the input image. During the denoising process, we follow (5) with $\sigma_t = 0$ for $t \in \{L, ..., 1\}$. We apply classifier guidance as introduced in [8] to lead the image generation to the desired healthy class h. For this, we pretrain a classifier network C on the noisy images x_t for $t \in \{1, ..., T\}$, to predict the class label of x. During the denoising process, the scaled gradient $s\nabla_{x_t} \log C(h|x_t, t)$ of the classifier is used to update $\epsilon_\theta(x_t, t)$. This iterative noising and denoising scheme is presented in Algorithm 1. We generate an image x_0 of the desired class h that preserves the basic structure of x. The anomaly map is then defined by the difference between x and x_0. The choice of the noise level L and the gradient scale s is crucial for the trade-off between detail-preserving image reconstruction and freedom for translation to a healthy subject.

Algorithm 1 Anomaly detection using noise encoding and classifier guidance

Input: input image x, healthy class label h, gradient scale s, noise level L
Output: synthetic image x_0, anomaly map a
for all t from 0 to $L-1$ **do**
$x_{t+1} \leftarrow x_t + \sqrt{\bar{\alpha}_{t+1}} \left[\left(\sqrt{\frac{1}{\bar{\alpha}_t}} - \sqrt{\frac{1}{\bar{\alpha}_{t+1}}} \right) x_t + \left(\sqrt{\frac{1}{\bar{\alpha}_{t+1}} - 1} - \sqrt{\frac{1}{\bar{\alpha}_t} - 1} \right) \epsilon_\theta(x_t, t) \right]$
end for
for all t from L to 1 **do**
$\hat{\epsilon} \leftarrow \epsilon_\theta(x_t, t) - s\sqrt{1-\bar{\alpha}_t} \nabla_{x_t} \log C(h|x_t, t)$
$x_{t-1} \leftarrow \sqrt{\bar{\alpha}_{t-1}} \left(\frac{x_t - \sqrt{1-\bar{\alpha}_t}\hat{\epsilon}}{\sqrt{\bar{\alpha}_t}} \right) + \sqrt{1-\bar{\alpha}_{t-1}}\hat{\epsilon}$
end for
$a \leftarrow \sum_{channels} |x - x_0|$
return x_0, a

3 Experiments

The DDPM is trained as proposed in [17] without data augmentation. We choose the hyperparameters for the DDPM model as described in the appendix of [8], for $T = 1000$ sampling steps. The model is trained with the Adam optimizer and the hybrid loss objective described in [17], with a learning rate of 10^{-4}, and a batch size of 10. By choosing the number of channels in the first layer as 128, and using one attention head at resolution 16, the total number of parameters is 113,681,160 for the diffusion model and 5,452,962 for the classifier. We train the class-conditional DDPM model for 50,000 iterations and the classifier network

for 20,000 iterations, which takes about one day on an NVIDIA Quadro RTX 6000 GPU. We used Pytorch 1.7.1 as software framework. The CheXpert and the BRATS2020 dataset are used for the evaluation of our method.

CheXpert This dataset contains lung X-ray images. For training, we choose 14,179 subjects of the healthy control group, as well as 16,776 subjects suffering from pleural effusions. The images are of size 256×256 and normalized to values between 0 and 1. The test set comprises 200 images of each class.

BRATS2020 This dataset contains 3D brain Magnetic Resonance (MR) images of subjects with a brain tumor, as well as pixel-wise ground truth labels. Every subject is scanned with four different MR sequences, namely, T1-weighted, T2-weighted, FLAIR, and T1-weighted with contrast enhancement. Since we focus on a 2D approach, we only consider axial slices. Each slice contains the aforementioned four channels, is padded to a size of 256×256, and normalized to values between 0 and 1. Since tumors mostly occur in the middle of the brain, we exclude the lowest 80 slices and the uppermost 26 slices. A slice is considered healthy if no tumor is found on the ground truth label mask. All other slices get the image-level label *diseased*. Our training set includes 5,598 healthy slices, and 10,607 diseased slices. The test set consists of 1,082 slices containing a tumor, and 705 slices without.

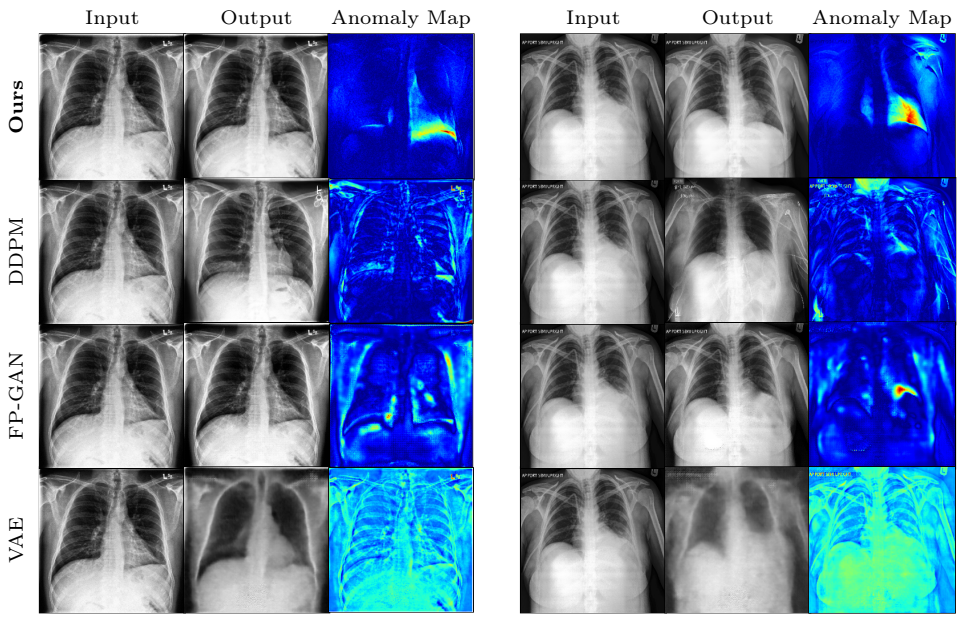


Fig. 2. Results for two X-ray images of the CheXpert dataset for $L = 500$ and $s = 100$.

4 Results and Discussion

For the evaluation of our method, we compare our method to the Fixed-Point GAN (FP-GAN) [24], and the variational autoencoder (VAE) proposed in [6]. As an ablation study, we add random noise for L steps to the input image using and perform the sampling using the DDPM sampling scheme with classifier guidance. In all experiments, we set $s = 100$ and $L = 500$. In Figure 2, we show two exemplary patient images of the CheXpert dataset, and apply all comparing methods to generate the corresponding healthy image. We observe that compared to the other methods, our approach generates realistic looking images and preserves all the details of the input image, which leads to a very detailed anomaly map. The other methods either change other parts image, or are not able to find an anomaly. Figure 3 shows the results for all four MR sequences for an exemplary image of the BRATS2020 dataset. More examples can be found in

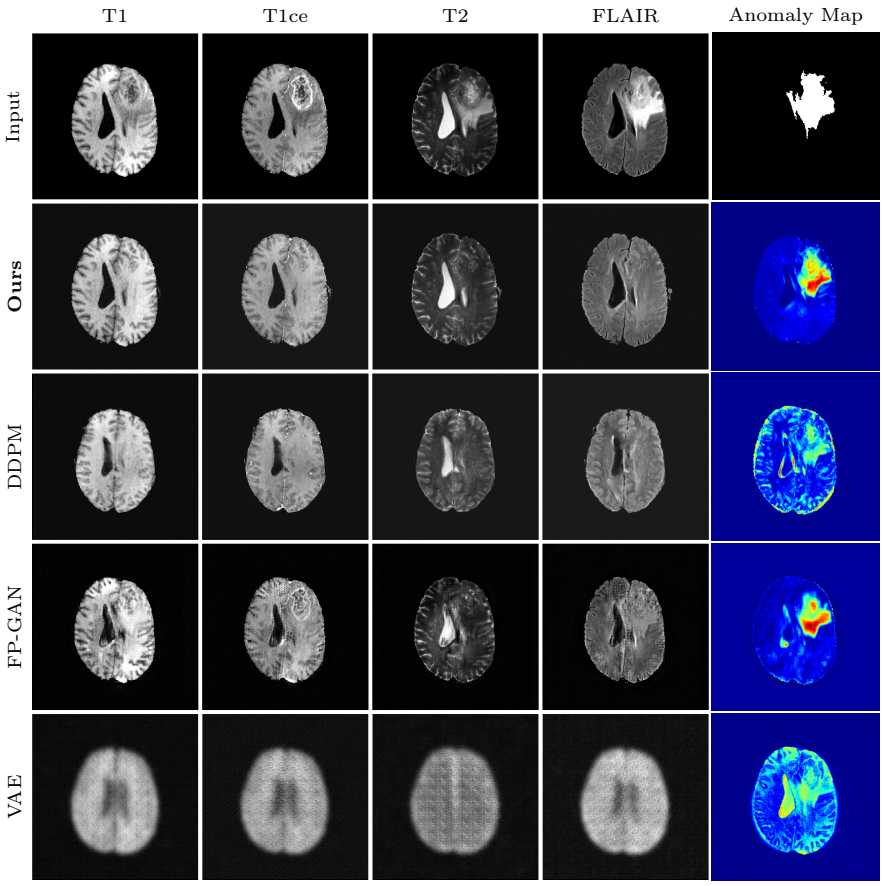


Fig. 3. Results for an image of the BRATS2020 dataset for $L = 500$ and $s = 100$.

the supplementary material. Of all methods, only the VAE tries to reconstruct the right ventricle. Comparing our results to the results of DDPM, we see that encoding information in noise using the deterministic noising process of DDIM brings the advantage that all details of the input image can be reconstructed. In contrast, we see that sampling with the DDPM approach changes the basic anatomy of the input image. The computation of a complete image translation takes about 158s. This longish running time is mainly due to the iterative image generation process. We could speed up this process by choosing a smaller L, or by skipping timesteps in the DDIM sampling scheme. However, we observed that this degrades the image quality.
In [15], using the reconstruction error as anomaly score has received some criticism. It was shown that a simple method based on histogram equalization could outperform neural networks and state that reconstruction quality does not correlate well with the Dice score. As an alternative, anomaly scores of other types of methods, i.e., the log-likelihood of density estimation models, will be explored in future work.

Hyperparameter Sensitivity Our method has two major hyperparameters, the classifier gradient scale s and the noise level L. We performed experiments to evaluate the sensitivity of our method to changes of s and L. On the BRATS2020 dataset, we have pixel-wise ground truth labels, which enable us to calculate the Dice score and the Area under the receiver operating statistics (AUROC) for diseased slices. For the Dice score, we use the average Otsu thresholding [18] on the anomaly maps. In Figure 4, we show the average Dice and AUROC scores on the test set with respect to the gradient scale s for different noise levels L. The scores for the comparing methods FP-GAN and VAE are shown in horizontal bars in Figure 4.
Figure 5 shows an exemplary FLAIR image. We fix $L = 500$ and show the sampled results for various values of s. If we choose s too small, the tumor cannot be removed. However, if we choose s too large, additional artefacts are introduced

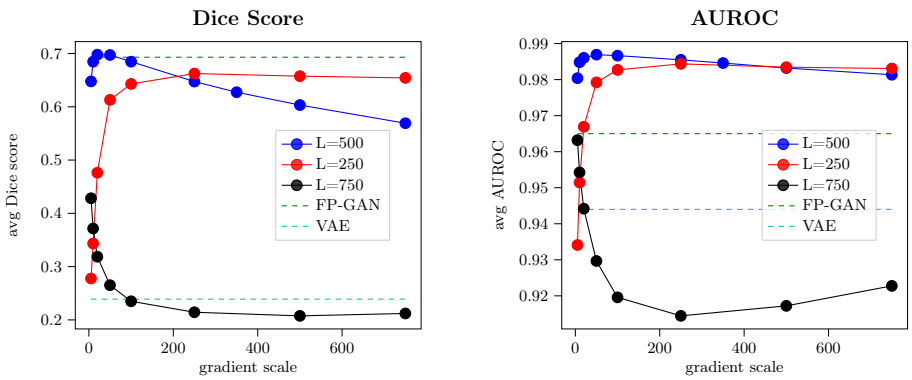


Fig. 4. Average Dice and AUROC scores on the test set for different s and L.

to the image. Those artefacts are mainly at the border of the brain, and lead to a decrease in the Dice score. In Figure 6, we fix $s = 100$, and show the sampled results for the same image for varying noise levels L. If L is chosen too large, this results in a destruction of the images. If L is chosen too small, the model does not have enough freedom to remove the tumor from the image.

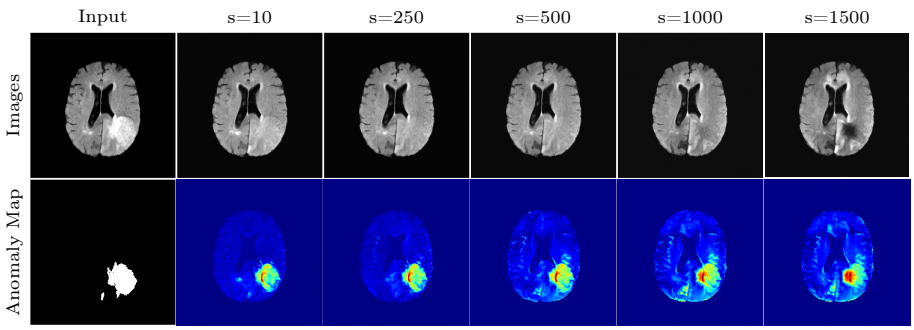


Fig. 5. Illustration of the effect of the gradient scale s for a fixed noise level $L = 500$.

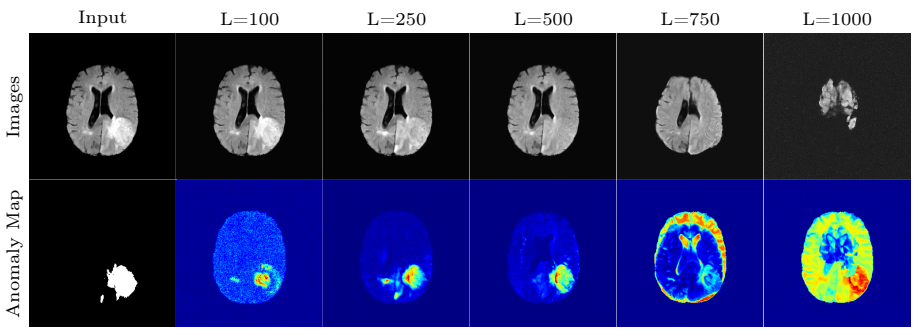


Fig. 6. Illustration of the effect of the noise level L, for a fixed gradient scale $s = 100$.

Translation of a Healthy Subject If an input image shows a healthy subject, our method should not make any changes to this image. In Figure 7, we evaluate our approach on a healthy slice of the BRATS dataset. We get a very detailed reconstruction of the image, resulting in an anomaly map close to zero.

5 Conclusion

In this paper, we presented a novel weakly supervised anomaly detection method by combining the iterative DDIM noising and denoising schemes, and classifier

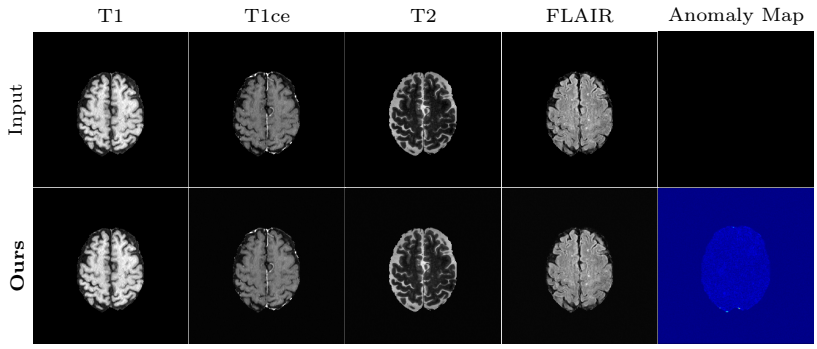


Fig. 7. Results of the presented method for an image without a tumor. The difference between the input image and the synthetic image is close to zero.

guidance. No changes were made to the loss function or the training scheme of the original implementations, making the training on other datasets straightforward. We applied our method for anomaly detection on two different medical datasets and successfully translated images of patients to images without pathologies. Our method only performs changes in the anomalous regions of the image to achieve the translation to a healthy subject. This improves the quality of the anomaly maps. We point out that we achieve a detail-consistent image-to-image translation without the need of changing the architecture or training procedure. We achieve excellent results on the BRATS2020 and the CheXpert dataset.

Acknowledgements This research was supported by the Novartis FreeNovation initiative and the Uniscientia Foundation (project #147-2018).

References

1. Arun, N.T., Gaw, N., Singh, P., Chang, K., Hoebel, K.V., Patel, J., Gidwani, M., Kalpathy-Cramer, J.: Assessing the validity of saliency maps for abnormality localization in medical imaging. arXiv preprint arXiv:2006.00063 (2020)
2. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C.: Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. Scientific data **4**(1), 1–13 (2017)
3. Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., Shinohara, R.T., Berger, C., Ha, S.M., Rozycki, M., et al.: Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. arXiv preprint arXiv:1811.02629 (2018)
4. Baranchuk, D., Voynov, A., Rubachev, I., Khrulkov, V., Babenko, A.: Label-efficient semantic segmentation with diffusion models. In: International Conference on Learning Representations (2022)

5. Baumgartner, C.F., Koch, L.M., Tezcan, K.C., Ang, J.X., Konukoglu, E.: Visual feature attribution using wasserstein gans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8309–8319 (2018)
6. Chen, X., Konukoglu, E.: Unsupervised detection of lesions in brain mri using constrained adversarial auto-encoders. arXiv preprint arXiv:1806.04972 (2018)
7. Choi, J., Kim, S., Jeong, Y., Gwon, Y., Yoon, S.: Ilvr: Conditioning method for denoising diffusion probabilistic models. arXiv preprint arXiv:2108.02938 (2021)
8. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in Neural Information Processing Systems **34** (2021)
9. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
10. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020)
11. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 590–597 (2019)
12. Kim, B., Han, I., Ye, J.C.: Diffusemorph: Unsupervised deformable image registration along continuous trajectory using diffusion models. arXiv preprint arXiv:2112.05149 (2021)
13. Kingma, D.P., Welling, M.: An introduction to variational autoencoders. arXiv preprint arXiv:1906.02691 (2019)
14. Marimont, S.N., Tarroni, G.: Anomaly detection through latent space restoration using vector quantized variational autoencoders. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). pp. 1764–1767. IEEE (2021)
15. Meissen, F., Kaissis, G., Rueckert, D.: Challenging current semi-supervised anomaly segmentation methods for brain mri. arXiv preprint arXiv:2109.06023 (2021)
16. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (BRATS). IEEE transactions on medical imaging **34**(10), 1993–2024 (2014)
17. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: Proceedings of the 38th International Conference on Machine Learning. vol. 139, pp. 8162–8171. PMLR (2021)
18. Otsu, N.: A threshold selection method from gray-level histograms. IEEE transactions on systems, man, and cybernetics **9**(1), 62–66 (1979)
19. Panwar, H., Gupta, P., Siddiqui, M.K., Morales-Menendez, R., Bhardwaj, P., Singh, V.: A deep learning and grad-cam based color visualization approach for fast detection of covid-19 cases using chest x-ray and ct-scan images. Chaos, Solitons & Fractals **140**, 110190 (2020)
20. Pinaya, W.H.L., Tudosiu, P.D., Gray, R., Rees, G., Nachev, P., Ourselin, S., Cardoso, M.J.: Unsupervised brain anomaly detection and segmentation with transformers. arXiv preprint arXiv:2102.11650 (2021)
21. Pirnay, J., Chai, K.: Inpainting transformer for anomaly detection. arXiv preprint arXiv:2104.13897 (2021)
22. Saharia, C., Chan, W., Chang, H., Lee, C.A., Ho, J., Salimans, T., Fleet, D.J., Norouzi, M.: Palette: Image-to-image diffusion models. arXiv preprint arXiv:2111.05826 (2021)

23. Sasaki, H., Willcocks, C.G., Breckon, T.P.: Unit-ddpm: Unpaired image translation with denoising diffusion probabilistic models. arXiv preprint arXiv:2104.05358 (2021)
24. Siddiquee, M.M.R., Zhou, Z., Tajbakhsh, N., Feng, R., Gotway, M.B., Bengio, Y., Liang, J.: Learning fixed points in generative adversarial networks: From image-to-image translation to disease detection and localization. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 191–200 (2019)
25. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
26. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
27. Wolleb, J., Sandkühler, R., Cattin, P.C.: Descargan: Disease-specific anomaly detection with weak supervision. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 14–24. Springer (2020)
28. Yang, J., Xu, R., Qi, Z., Shi, Y.: Visual anomaly detection for images: A survey. arXiv preprint arXiv:2109.13157 (2021)
29. Zhou, C., Paffenroth, R.C.: Anomaly detection with robust deep autoencoders. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 665–674 (2017)
30. Zimmerer, D., Kohl, S.A., Petersen, J., Isensee, F., Maier-Hein, K.H.: Context-encoding variational autoencoder for unsupervised anomaly detection. arXiv preprint arXiv:1812.05941 (2018)

Supplementary Material

1 Additional Results on the CheXpert Dataset

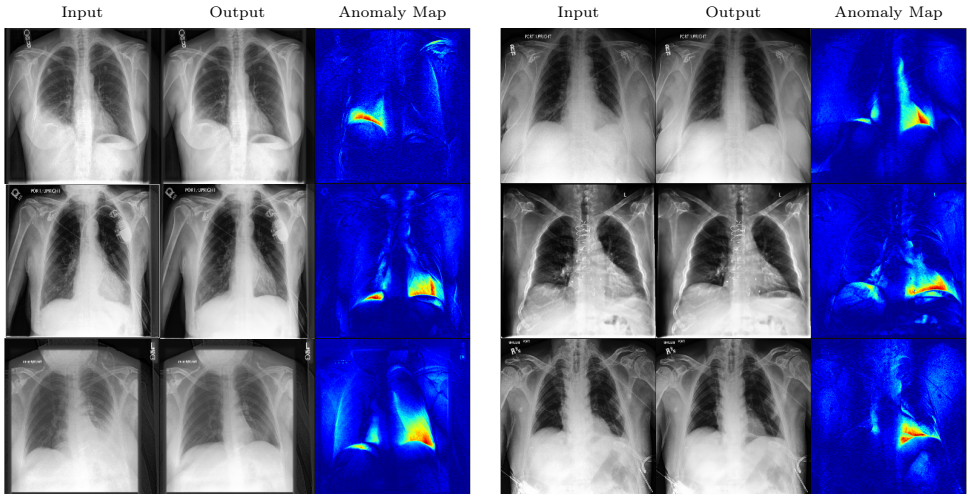


Fig. 1. Additional results of our method for diseased subjects, for $L = 500$ and $s = 100$.

2 Additional Results on the Brats2020 Dataset

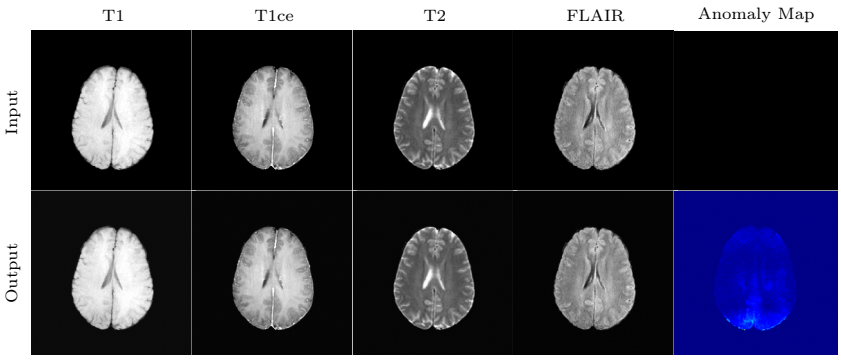


Fig. 2. Results of our method for a healthy subject, for $L = 500$ and $s = 100$.

2

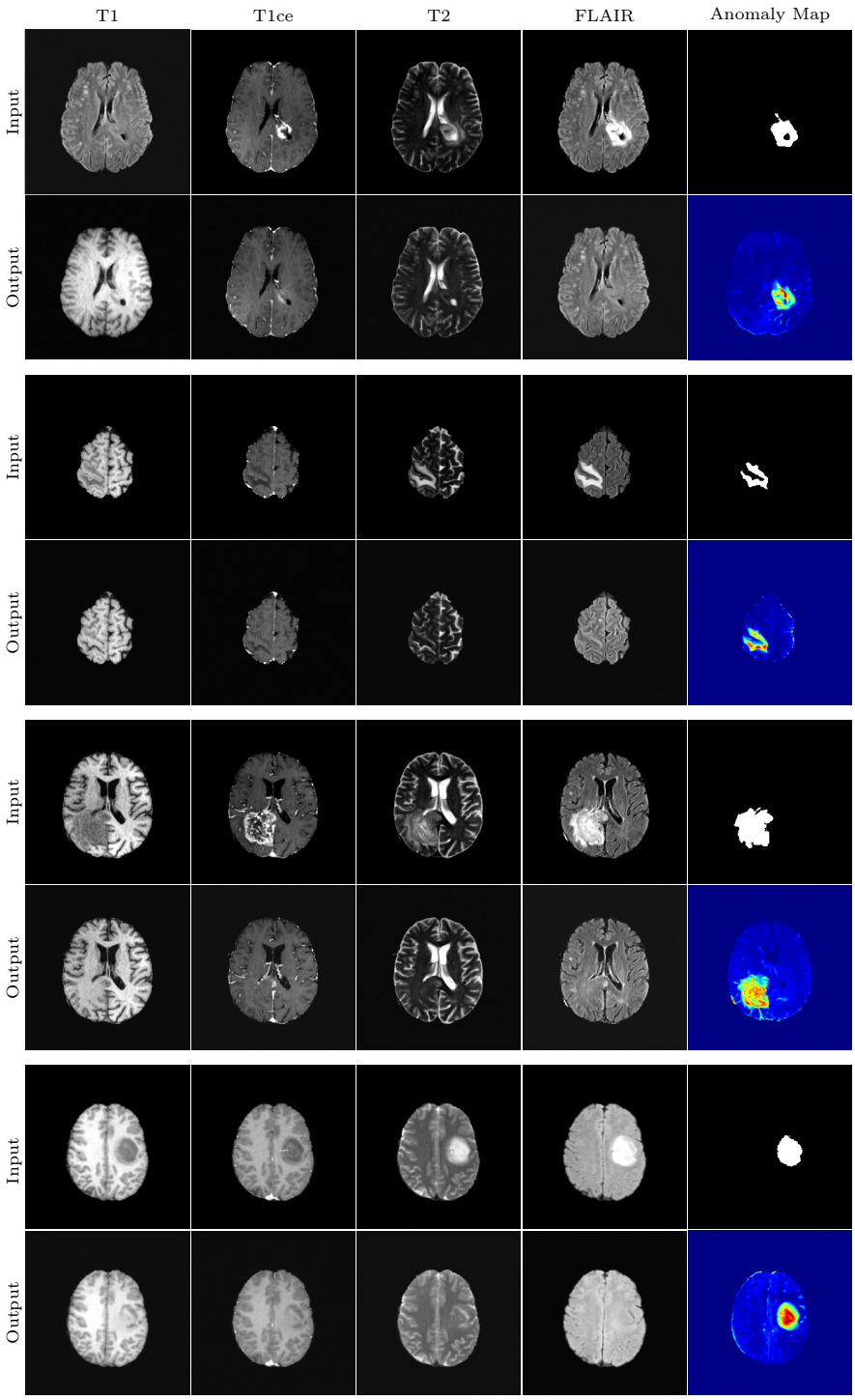


Fig. 3. Additional results of our method for diseased subjects, for $L = 500$ and $s = 100$.

Chapter 8

The Swiss Army Knife for Image-to-Image Translation: Multi-Task Diffusion Models

In Chapter 7, image-to-image translation is performed based on a binary classification model that distinguishes between healthy and diseased subjects. The following manuscript presents an extension of this idea, and we perform image-to-image translation based on a regression task and a segmentation task. We train a denoising diffusion model and a separate task-specific model on the same dataset. By adding the gradient of the task-specific model, we demonstrate that image-to-image translation using denoising diffusion implicit models and gradient guidance leads to an output image with the desired characteristics. Other features such as the background are preserved in great detail. We evaluate our method with an age regression task on facial photos, simulate tumor growth and inpaint tumors in brain MR images of healthy subjects at a desired location. The proposed framework is flexible and can be applied to various tasks without the need to retrain the diffusion model.
This technical report was written in joint first authorship with Robin Sandkühler, who had the initial idea for this approach. The method was developed based on an equal discussion, and the code implementation and the experiments were performed by Julia Wolleb.

Technical Report This manuscript has been submitted to arxiv.org[1].

[1] https://arxiv.org/pdf/2204.02641v1.pdf

The Swiss Army Knife for Image-to-Image Translation: Multi-Task Diffusion Models

Julia Wolleb*, Robin Sandkühler*, Florentin Bieder, Philippe C. Cattin

Department of Biomedical Engineering, University of Basel, Allschwil, Switzerland
julia.wolleb@unibas.ch

Abstract. Recently, diffusion models were applied to a wide range of image analysis tasks. We build on a method for image-to-image translation using denoising diffusion implicit models and include a regression problem and a segmentation problem for guiding the image generation to the desired output. The main advantage of our approach is that the guidance during the denoising process is done by an external gradient. Consequently, the diffusion model does not need to be retrained for the different tasks on the same dataset. We apply our method to simulate the aging process on facial photos using a regression task, as well as on a brain magnetic resonance (MR) imaging dataset for the simulation of brain tumor growth. Furthermore, we use a segmentation model to inpaint tumors at the desired location in healthy slices of brain MR images. We achieve convincing results for all problems.

Keywords: Diffusion models · Image-to-image translation · Regression · Segmentation.

1 Introduction

For many applications, it is of great interest to perform image-to-image translation such that the output image is changed in some regions to the desired characteristics and unchanged in other regions, e.g., the image background. Most approaches rely on

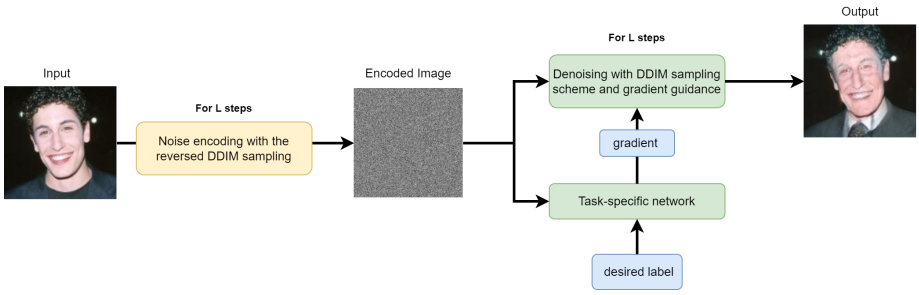


Fig. 1. Proposed scheme for image-to-image translation for the example of age regression.

* equal contribution

generative adversarial nets (GANs) [8] or variational autoencoders [17]. However, the adversarial training of GANs can be difficult and requires a lot of hyperparameter tuning. Furthermore, a big challenge is that only image features related to the desired output characteristics should be changed. To circumvent those issues, we propose an approach based on Denoising Diffusion Probabilistic Models (DDPMs)[9] and the sampling scheme of Denoising Diffusion Implicit Models (DDIMs) [24].

In [26], we performed image-to-image translation between different classes by training a DDPM, and using the DDIM noising and denoising scheme and classifier guidance during sampling. We build on this approach and present a method for image-to-image translation for variable tasks. Figure 1 shows an overview of the proposed approach. First, we separately train a DDPM as well as an external task-specific model on the same dataset. In this work, this external model is a regression or a segmentation model. The case of a classification model is already described in [26]. Image-to-image translation is performed only during the sampling process. For this, we first encode the information about the input image with the iterative noising process of DDIMs. During the denoising process using the DDIM sampling scheme, we inject the gradient of the task-specific model in each sampling step. By scaling this gradient, the generation of the output image is guided towards the desired output.

For the regression problem, we apply our algorithm to a dataset of facial images [7] for age prediction, where image generation is guided towards a desired age. An example for this is given in Figure 2, where we translate photos of two exemplary people of age 40 to photos showing the same person at various other ages. Moreover, we use a regression task on the BRATS2020 challenge [15,1,2] for the simulation of brain tumor growth. The segmentation problem is applied to the BRATS2020 dataset for the translation of images of healthy subjects to images containing a brain tumor at a location that we can freely choose.

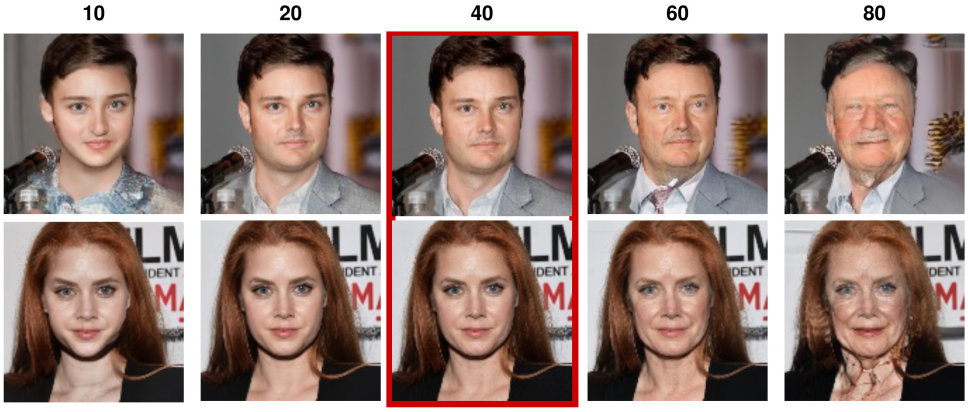


Fig. 2. Results of our method for the age regression task on facial photos. The original images framed in red show exemplary subjects of age 40. With our approach, they are translated to images matching a desired age value $i \in \{10, 20, 60, 80\}$.

Related Work In computer vision, image-to-image translation towards specific image attributes has been a task of great interest and includes style transfer [14], relighting or colorization tasks [21], and changing facial attributes such as hair color or gender [5]. The translation of images to subjects of another age has been explored by using autoencoders [12] based on facial landmarks, or by adapting GANs [10,23]. A big challenge of those approaches is that only task-related features should be changed, and the rest of the image remains consistent with the input image.
Lately, DDPMs were in focus due to their success in tasks such as image generation [6], image-to-image translation [22,4], segmentation [3,27], reconstruction [20] and registration[11]. In [19], denoising diffusion models were used for multivariate probabilistic time series forecasting. In [13], the gradient of image-text or image matching scores are used to guide the generation of synthetic images.
The presented method is based on our previous work [26], where image-to-image translation was performed based on a binary classification problem. This approach used classifier guidance during the sampling process, such that the training of the DDPM is not changed and a pretrained model can be used. The big advantages are the straightforward training process, and the fact that only features related to the classification problem are changed. The rest of the image is preserved.
In [18], diffusion autoencoders were proposed for meaningful representation learning. They encode image information in a latent space and use a conditional DDIM to manipulate the image attributes. Very recently, [25] proposed the training of two diffusion models for translations between arbitrary pairs of source-target domains. In contrast to our method, those approaches depend on a manipulation of the latent space, and the diffusion models need to be retrained for each application. For our method, the same diffusion model can be used for various applications, as the guidance towards the desired attributes is done only during the sampling process by an external gradient.

2 Method

The goal is to perform image-to-image translation such that the output image matches task-specific criteria. Thereby, it is important that only features related to the task are changed, and all other features are preserved. Our method follows [26], where the task-specific criteria is given by a binary classifier. In this work, we adapt this method with a regression model as described in Algorithm 1, as well as with a segmentation model, as described in Algorithm 2.
We implement a DDPM according to [9,16]. For an input image x, we add small amounts of noise for many steps T, such that we get a series of increasingly noisy images $\{x_0, x_1, ..., x_T\}$. A diffusion model is given by a U-Net ϵ_θ, which is trained with the MSE loss to predict x_{t-1} from x_t. During sampling, a synthetic image x_0 can be generated from $x_T \sim \mathcal{N}(0, \mathbf{I})$ by predicting x_{t-1} from x_t for $t \in \{T, ..., 1\}$ using (3). During training, we can explicitly write the forward noising process as

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, \quad \text{with } \epsilon \sim \mathcal{N}(0, \mathbf{I}). \tag{1}$$

Here, we define $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$, where $\beta_1, ..., \beta_T$ denote the forward process variances. This noisy image x_t given in (1) serves as input for the U-Net ϵ_θ,

which is trained using the MSE loss

$$\mathcal{L} := ||\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)||_2^2, \quad \text{with } \epsilon \sim \mathcal{N}(0, \mathbf{I}). \tag{2}$$

We can then predict x_{t-1} from x_t with

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\left(\frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t,t)}{\sqrt{\bar{\alpha}_t}}\right) + \sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}\epsilon_\theta(x_t,t) + \sigma_t\epsilon. \tag{3}$$

In DDIMs, we set $\sigma_t = 0$, which results in a deterministic sampling process. Equation (3) can be interpreted as the Euler method to solve the ordinary differential equation (ODE) described in [24]. By solving the reversed ODE, we can reverse the generation process. Consequently, using enough discretization steps, we can encode x_{t+1} given x_t with

$$x_{t+1} = x_t + \sqrt{\bar{\alpha}_{t+1}}\left[\left(\sqrt{\frac{1}{\bar{\alpha}_t}} - \sqrt{\frac{1}{\bar{\alpha}_{t+1}}}\right)x_t + \left(\sqrt{\frac{1}{\bar{\alpha}_{t+1}}-1} - \sqrt{\frac{1}{\bar{\alpha}_t}-1}\right)\epsilon_\theta(x_t,t)\right]. \tag{4}$$

For image-to-image translation, we define a noise level $L \in \{1, ..., T\}$, as proposed in [26]. By applying (4) for $t \in \{0, ..., L-1\}$, we can encode an image x_0 in a noisy image x_L. With this deterministic, iterative noising process, the information of the input image x is stored in x_L.
In addition to the diffusion model, we train a separate task-specific network on our set of noisy images $\{x_0, x_1, ..., x_T\}$. During the denoising process, we follow (3) with $\sigma_t = 0$ for $t \in \{L, ..., 1\}$, and add the gradient of the task-specific network in every step during sampling to lead the image generation to the desired output characteristics. The case of classification was already presented in [26]. In Sections 2.1 and 2.2, we investigate the tasks of regression and segmentation.

2.1 Regression

The regression model R follows the architecture of the encoder of the diffusion model, and is trained with the MSE loss. The iterative noising and denoising scheme for image-to-image translation for a regression problem is presented in Algorithm 1. We encode an image x in a noisy image x_L, and define a desired value i. During the denoising process, the gradient $\nabla_{x_t}R(x_t,t)$ of the regression model is used to update $\epsilon_\theta(x_t,t)$. The sign of the gradient defines the direction in which the generation process is influenced, e.g., whether the subject is made younger or older.
To guide the image generation to the desired value i for the regression task, we define the gradient scale $s_t = i - R(x_t,t)$, such that the influence of the gradient gets smaller if the predicted value is close to the desired value. If the predicted value surpasses i, the sign of s_t is changed, and the denoising process is guided back in the other direction. An additional gradient scale c is constant over time and can be used to further amplify the gradient, as we explored in [26].

Algorithm 1 Regression guidance

Input: input image x, desired value i, noise level L, constant gradient scale c
Output: synthetic image x_0
$x_0 \leftarrow x$
for all t from 0 to $L-1$ **do**
$x_{t+1} \leftarrow x_t + \sqrt{\bar{\alpha}_{t+1}} \left[\left(\sqrt{\frac{1}{\bar{\alpha}_t}} - \sqrt{\frac{1}{\bar{\alpha}_{t+1}}} \right) x_t + \left(\sqrt{\frac{1}{\bar{\alpha}_{t+1}} - 1} - \sqrt{\frac{1}{\bar{\alpha}_t} - 1} \right) \epsilon_\theta(x_t, t) \right]$
end for
for all t from L to 1 **do**
$s_t \leftarrow i - R(x_t, t)$
$\hat{\epsilon} \leftarrow \epsilon_\theta(x_t, t) - s_t c \sqrt{1 - \bar{\alpha}_t} \nabla_{x_t} R(x_t, t)$
$x_{t-1} \leftarrow \sqrt{\bar{\alpha}_{t-1}} \left(\frac{x_t - \sqrt{1-\bar{\alpha}_t}\hat{\epsilon}}{\sqrt{\bar{\alpha}_t}} \right) + \sqrt{1 - \bar{\alpha}_{t-1}} \hat{\epsilon}$
end for
return x_0

2.2 Segmentation

The algorithm for segmentation guidance can be found in Algorithm 2. The iterative noise encoding scheme stays the same as in Section 2.1. However, instead of a regression network, we train a segmentation network S on the noisy images $\{x_0, x_1, ..., x_T\}$ with the cross-entropy loss. The architecture of the segmentation model follows the U-Net architecture of the diffusion model. We further define a desired label mask z, and compute the binary cross-entropy loss H between the output of the segmentation network $S(x_t, t)$ and a desired label mask z. We can then guide the image generation towards an image that matches z by using the gradient $\nabla_{x_t} H$ during the denoising process. As already proposed in Section 2.1, a constant gradient scale c can be applied to amplify this gradient.

Algorithm 2 Segmentation guidance

Input: input image x, desired label mask z, noise level L, constant gradient scale c, number of pixels P
Output: synthetic image x_0
$x_0 \leftarrow x$
for all t from 0 to $L-1$ **do**
$x_{t+1} \leftarrow x_t + \sqrt{\bar{\alpha}_{t+1}} \left[\left(\sqrt{\frac{1}{\bar{\alpha}_t}} - \sqrt{\frac{1}{\bar{\alpha}_{t+1}}} \right) x_t + \left(\sqrt{\frac{1}{\bar{\alpha}_{t+1}} - 1} - \sqrt{\frac{1}{\bar{\alpha}_t} - 1} \right) \epsilon_\theta(x_t, t) \right]$
end for
for all t from L to 1 **do**
$H \leftarrow -\frac{1}{P} \sum_{j=1}^{P} \left(z_j \log S(x_t, t)_j + (1 - z_j) \log(1 - S(x_t, t)_j) \right)$
$\hat{\epsilon} \leftarrow \epsilon_\theta(x_t, t) - c \sqrt{1 - \bar{\alpha}_t} \nabla_{x_t} H$
$x_{t-1} \leftarrow \sqrt{\bar{\alpha}_{t-1}} \left(\frac{x_t - \sqrt{1-\bar{\alpha}_t}\hat{\epsilon}}{\sqrt{\bar{\alpha}_t}} \right) + \sqrt{1 - \bar{\alpha}_{t-1}} \hat{\epsilon}$
end for
return x_0

3 Experiments

The DDPM is trained as proposed in [16] without any data augmentation. We choose $T = 1000$, $L = 400$. The other hyperparameters for the DDPM are described in the appendix of [6]. The model is trained with the Adam optimizer and the hybrid loss objective described in [16], with a learning rate of 10^{-4}, and a batch size of 10. The number of channels in the first layer is chosen as 128, and using one attention head at resolution 16. The regression model has a depth of 4 and uses attention heads at the resolution of $8, 16$ and 32. We set the number of channels in the first layer to 32 for the BRATS2020 dataset, and to 128 for the dataset of facial photos. Training was performed on an NVIDIA Quadro RTX 6000 GPU, with Pytorch 1.7.1 as software framework.

Facial Photos The dataset of facial photos shows people aged between 1 and 100. The training dataset comprises 185,631 images of size $3 \times 128 \times 128$. The test set includes 47,568 images. All images are normalized to values between 0 and 1. The diffusion model is trained for 500,000 iterations, due to the high variability in the data. The regression model is trained for 80,000 iterations. The number of parameters is 85,606,150 for the diffusion model, and 67,061,121 for the regression model. Image-to-image translation for one image takes 73 s for the age regression task.

BRATS2020 The BRATS2020 dataset of 3D brain Magnetic Resonance (MR) images of subjects with a brain tumor is described in [26]. We only consider 2D axial slices. For each slice, four different MR sequences as well as the pixel-wise ground truth segmentation of the tumor is given. The images are of size $4 \times 256 \times 256$, where each channel shows one of the four MR sequences, namely T1-weighted, T2-weighted, FLAIR, and T1-weighted with contrast enhancement (T1ce). All images are normalized to values between 0 and 1. For the regression problem, the relative tumor size is calculated as the ratio between the tumor size and the brain size for each slice. Our training set includes 16,205 slices, whereas there are 1,787 images in the test set. We train the diffusion model for 50,000 iterations, the regression and segmentation models for 20,000 iterations. The number of model parameters is 113,681,160 for the diffusion and the segmentation model, and 5,452,833 for the regression model. Image-to-image translation for one image takes 84 s for the regression task and 102 s for the segmentation task.

4 Results and Discussion

4.1 Age Regression on Facial Photos

In Figure 3, we present exemplary images of the test set, as well as the output of our model, if we set $s_t = 1 \quad \forall t$. This results in aging of the subjects. In Figure 4, we present further examples, as well as the output of our model, if we set $s_t = -1 \quad \forall t$. We see that the output images show younger subjects, whereas the background and other features such as hair and clothes are preserved. For this dataset, we choose $c = 5$.
As described in Section 2.1, we set $s_t = i - R(x_t, t)$ if we wish to generate an image of the desired age i. In Figure 2, we show the aging process for two subjects of age 40, and set the desired age values to $i \in \{10, 20, 60, 80\}$.

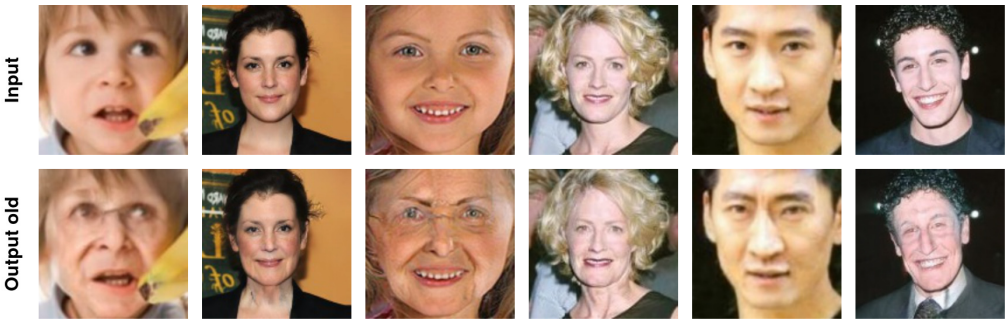


Fig. 3. Results of our method for $s_t = 1$. The positive gradient makes the subjects older.

Fig. 4. Results of our method for $s_t = -1$. The negative gradient makes the subjects younger.

4.2 Regression on the Relative Tumor Size

We train a regression model on the BRATS2020 dataset described in Section 3. The value for each slice is defined as the ratio of the area of the ground truth segmentation mask and the area of the brain. Like in Section 4.1, we can make the tumor grow or shrink by changing the sign of the gradient. In Figure 5, we present the input and output images for all four MR sequences for $s_t = 1$. The output shows images with an enlarged tumor. This can also be seen on the difference map in the last column, where the absolute difference between the input and the output image, summed over all 4 channels, is presented. On the other hand, Figure 6 shows the output of our method for $s_t = -1$, resulting in a smaller tumor. On the BRATS2020 dataset, we choose $c = 1000$.
Just like in the age regression problem, we can influence the tumor size by providing a desired value i. In Figure 7, we show the results for various desired values $i \in \{0, 0.05, 0.1, 0.2\}$, where the original value is 0.08. All four MR sequences as well as the absolute difference map between the input and the output image are provided.

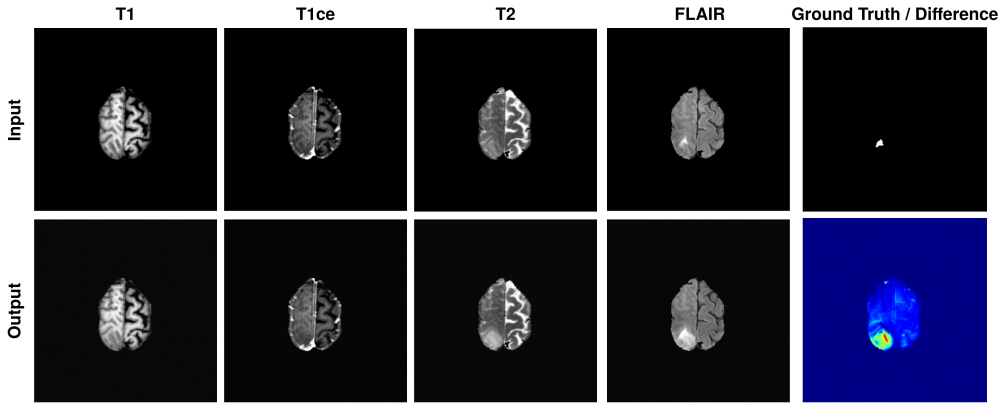


Fig. 5. Results of our method for $s_t = 1$, which results in an enlarged tumor.

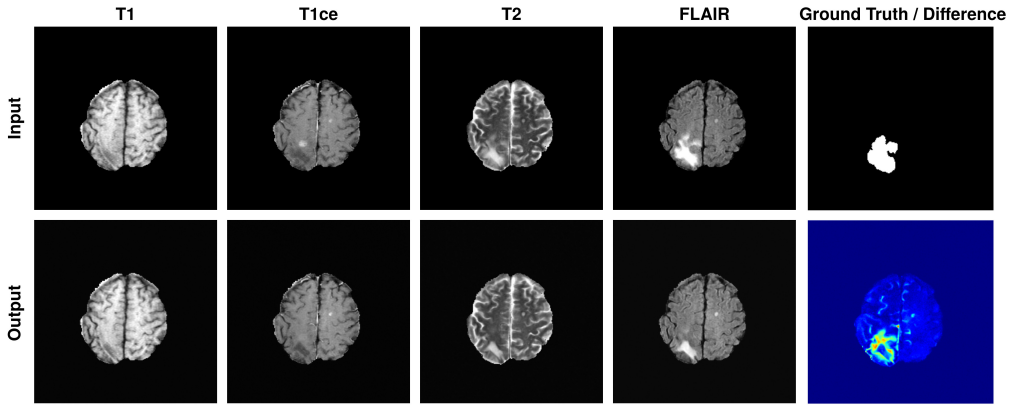


Fig. 6. Results of our method for $s_t = -1$, which leads to a smaller tumor.

4.3 Tumor Generation using Segmentation Models

We train a fully supervised segmentation model on the BRATS2020 dataset for brain tumor segmentation. For image-to-image translation, we aim to translate an image showing a healthy slice into a slice containing a tumor. For this, we define a pixel-wise label mask where we want the model to insert a tumor.
In the first row of Figure 8, we present the input image showing a healthy slice, as well as the desired label mask in red. The output image shows a brain MR image containing a fake tumor. By considering the difference between the input and output image, we see that this fake tumor was drawn in the desired area. For these experiments, we set $c = 5$. This approach could be helpful for the generation of artificial data for the training of anomaly detection methods.

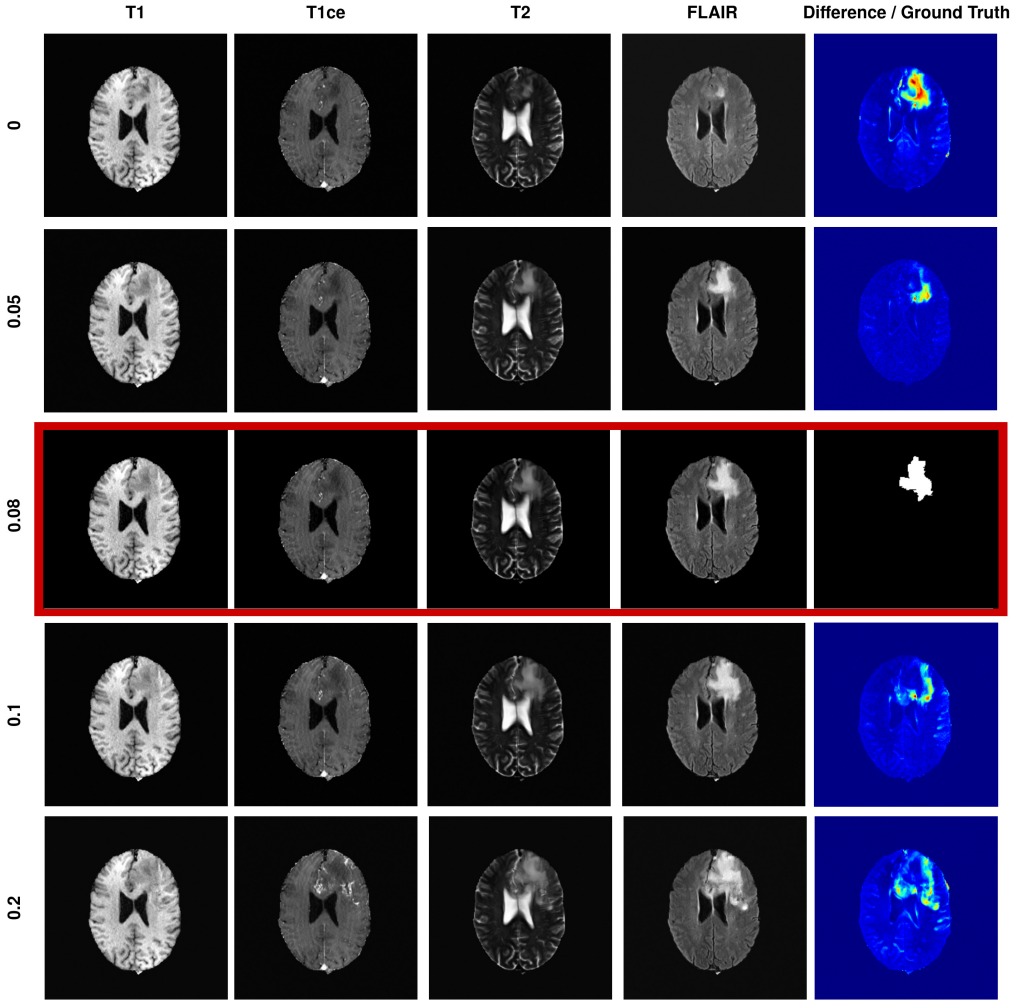


Fig. 7. Results of our method for $s_t = i - R(x_t, t)$ and $i \in \{0, 0.05, 0.1, 0.2\}$. The original image framed in red has a ratio of 0.08. The difference maps in the last column highlight the regions that are changed during image-to-image translation.

5 Conclusion

In this paper, we present an image-to-image translation method based on DDIMs for different specific tasks using gradient guidance. In addition to a classical DDPM, a separate network for the specific task is trained. We show that our method can translate images to an output matching a desired value or label mask of a regression or segmentation problem. The resulting images are only altered in features related to that desired characteristics, and the rest of the image is preserved. The big advantage of this approach

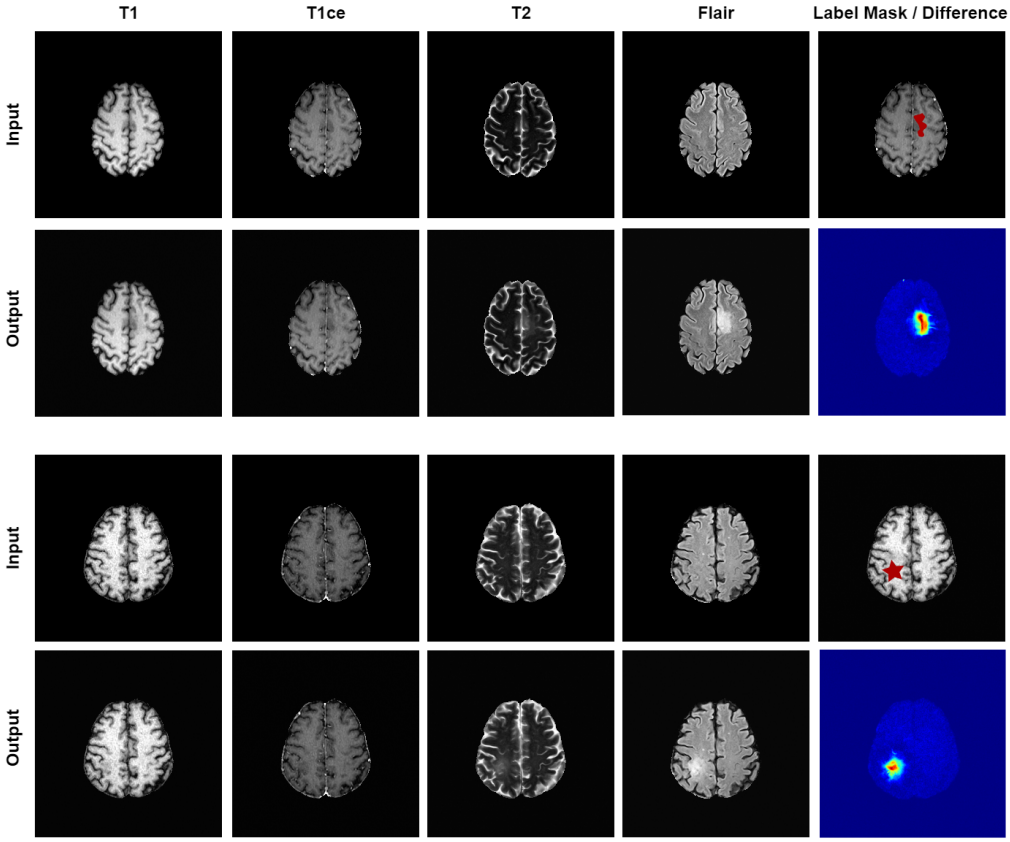


Fig. 8. Results for the segmentation task on the BRATS2020 dataset. The label mask is shown in the last column in red, together with one channel of the input image as a reference in the background. The difference map shows that a tumor was inpainted in the desired area.

is that the same diffusion model can be used for multiple tasks, i.e., classification, segmentation or regression.
We applied our method to a dataset of facial photos for age regression, and on the BRATS2020 dataset for brain tumor growth. By using the segmentation task, we are able to insert brain tumors in healthy slices of the BRATS2020 dataset. This can be useful for the generation of artificial training data for anomaly detection methods. We achieve convincing results on both datasets for both tasks. Future work includes speeding up the sampling process, and the extension to 3D data.

Acknowledgements This research was supported by the Novartis FreeNovation initiative and the Uniscientia Foundation (project #147-2018).

References

1. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C.: Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. Scientific data **4**(1), 1–13 (2017)
2. Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., Shinohara, R.T., Berger, C., Ha, S.M., Rozycki, M., et al.: Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. arXiv preprint arXiv:1811.02629 (2018)
3. Baranchuk, D., Rubachev, I., Voynov, A., Khrulkov, V., Babenko, A.: Label-efficient semantic segmentation with diffusion models. arXiv preprint arXiv:2112.03126 (2021)
4. Choi, J., Kim, S., Jeong, Y., Gwon, Y., Yoon, S.: Ilvr: Conditioning method for denoising diffusion probabilistic models. arXiv preprint arXiv:2108.02938 (2021)
5. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8789–8797 (2018)
6. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in Neural Information Processing Systems **34** (2021)
7. Frențescu, M.: Age prediction, `https://www.kaggle.com/datasets/mariafrenti/age-prediction`, retrieved 26.02.2022
8. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
9. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020)
10. Huang, Z., Chen, S., Zhang, J., Shan, H.: Pfa-gan: Progressive face aging with generative adversarial network. IEEE Transactions on Information Forensics and Security **16**, 2031–2045 (2020)
11. Kim, B., Han, I., Ye, J.C.: Diffusemorph: Unsupervised deformable image registration along continuous trajectory using diffusion models. arXiv preprint arXiv:2112.05149 (2021)
12. Lan, L.C., Liu, T.J., Liu, K.H.: Age regression with specific facial landmarks by dual discriminator adversarial autoencoder. In: 2021 IEEE International Conference on Image Processing (ICIP). pp. 2718–2722. IEEE (2021)
13. Liu, X., Park, D.H., Azadi, S., Zhang, G., Chopikyan, A., Hu, Y., Shi, H., Rohrbach, A., Darrell, T.: More control for free! image synthesis with semantic diffusion guidance. arXiv preprint arXiv:2112.05744 (2021)
14. Liu, Z.S., Kalogeiton, V., Cani, M.P.: Multiple style transfer via variational autoencoder. In: 2021 IEEE International Conference on Image Processing (ICIP). pp. 2413–2417. IEEE (2021)
15. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (BRATS). IEEE transactions on medical imaging **34**(10), 1993–2024 (2014)
16. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: Proceedings of the 38th International Conference on Machine Learning. vol. 139, pp. 8162–8171. PMLR (2021)
17. Pang, Y., Lin, J., Qin, T., Chen, Z.: Image-to-image translation: Methods and applications. IEEE Transactions on Multimedia (2021)
18. Preechakul, K., Chatthee, N., Wizadwongsa, S., Suwajanakorn, S.: Diffusion autoencoders: Toward a meaningful and decodable representation. arXiv preprint arXiv:2111.15640 (2021)

19. Rasul, K., Seward, C., Schuster, I., Vollgraf, R.: Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. In: International Conference on Machine Learning. pp. 8857–8868. PMLR (2021)
20. Saharia, C., Chan, W., Chang, H., Lee, C.A., Ho, J., Salimans, T., Fleet, D.J., Norouzi, M.: Palette: Image-to-image diffusion models. arXiv preprint arXiv:2111.05826 (2021)
21. Santhanam, V., Morariu, V.I., Davis, L.S.: Generalized deep image to image regression. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5609–5619 (2017)
22. Sasaki, H., Willcocks, C.G., Breckon, T.P.: Unit-ddpm: Unpaired image translation with denoising diffusion probabilistic models. arXiv preprint arXiv:2104.05358 (2021)
23. Sharma, N., Sharma, R., Jindal, N.: Prediction of face age progression with generative adversarial networks. Multimedia Tools and Applications **80**(25), 33911–33935 (2021)
24. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
25. Su, X., Song, J., Meng, C., Ermon, S.: Dual diffusion implicit bridges for image-to-image translation. arXiv preprint arXiv:2203.08382 (2022)
26. Wolleb, J., Bieder, F., Sandkühler, R., Cattin, P.C.: Diffusion models for medical anomaly detection. arXiv preprint arXiv:2203.04306 (2022)
27. Wolleb, J., Sandkühler, R., Bieder, F., Valmaggia, P., Cattin, P.C.: Diffusion models for implicit image segmentation ensembles. arXiv preprint arXiv:2112.03145 (2021)

Chapter 9

Learn to Ignore: Domain Adaptation for Multi-Site MRI Analysis

In this work, we focus on the problem of limited availability of large image datasets in medical applications. This is especially the case for MR data, where different MR scanners introduce a bias that limits the performance of machine learning models. We train a binary classification model on brain MR images to distinguish between healthy controls and MS patients. Such a classification network is needed in more advanced weakly supervised anomaly detection methods such as the ones presented in Chapters 6 and 7. Due to the use of different scanners and acquisition protocols, only a small dataset acquired with the same settings is at hand to train the model, leading to a poor generalization quality. To overcome this issue, we add data from other scanners. However, if an additional dataset does not include all classes of the task, the learning of the classification model can be biased to the device or place of acquisition.
In the following paper, we introduce specific additional constraints on the latent space, which can be included in any classification network. With this setup, the model learns to ignore the scanner-related features present in the images while learning disease-specific features relevant to the classification task. This work presents a step towards developing robust deep learning methods across different MR scanners.

Publication. The proposed approach is accepted at the *25th International Conference on Medical Image Computing and Computer-Assisted Intervention* (MICCAI), September 2022, Singapore. The manuscript was published as part of the conference proceedings [195]. The code for this framework is open-source[1].

[1] https://gitlab.com/cian.unibas.ch/L2I

Learn to Ignore: Domain Adaptation for Multi-Site MRI Analysis

Julia Wolleb[1], Robin Sandkühler[1], Florentin Bieder[1], Muhamed Barakovic[1], Nouchine Hadjikhani[3,4], Athina Papadopoulou[1,2], Özgür Yaldizli[1,2], Jens Kuhle[2], Cristina Granziera[1,2], and Philippe C. Cattin[1]

[1] Department of Biomedical Engineering, University of Basel, Allschwil, Switzerland
[2] University Hospital Basel, Switzerland
[3] Massachusetts General Hospital, Harvard Medical School, Charlestown MA, USA
[4] Gillberg Neuropsychiatry Center, Sahlgrenska Academy, University of Gothenburg, Sweden
julia.wolleb@unibas.ch

Abstract. The limited availability of large image datasets, mainly due to data privacy and differences in acquisition protocols or hardware, is a significant issue in the development of accurate and generalizable machine learning methods in medicine. This is especially the case for Magnetic Resonance (MR) images, where different MR scanners introduce a bias that limits the performance of a machine learning model. We present a novel method that learns to ignore the scanner-related features present in MR images, by introducing specific additional constraints on the latent space. We focus on a real-world classification scenario, where only a small dataset provides images of all classes. Our method *Learn to Ignore (L2I)* outperforms state-of-the-art domain adaptation methods on a multi-site MR dataset for a classification task between multiple sclerosis patients and healthy controls.

Keywords: domain adaptation, scanner bias, MRI

1 Introduction

Due to its high soft-tissue contrast, Magnetic Resonance Imaging (MRI) is a powerful diagnostic tool for many neurological disorders. However, compared to other imaging modalities like computed tomography, MR images only provide relative values for different tissue types. These relative values depend on the scanner manufacturer, the scan protocol, or even the software version. We refer to this problem as the scanner bias. While human medical experts can adapt to these relative changes, they represent a major problem for machine learning methods, leading to a low generalization quality of the model. By defining different scanner settings as different domains, we look at this problem from the perspective of domain adaptation (DA) [3], where the main task is learned on a *source domain*. The model then should perform well on a different *target domain*.

1.1 Related Work

An overview of DA in medical imaging can be found at [13]. One can generally distinguish between unsupervised domain adaptation (UDA) [1], where the target domain data is unlabeled, or supervised domain adaptation (SDA) [28], where the labels of the target domain are used during training.
The problem of scanner bias is widely known to disturb the automated analysis of MR images [22], and a lot of work already tackles the problem of multi-site MR harmonization [11]. Deepharmony [7] uses paired data to change the contrast of MRI from one scanner to another scanner with a modified U-Net. Generative Adversarial Networks aim to generate new images to overcome the domain shift [24]. These methods modify the intensities of each pixel before training for the main task. This approach is preferably avoided in medical applications, as it bears the risk of removing important pixel-level information required later for other tasks, such as segmentation or anomaly detection.
Domain-adversarial neural networks [12] can be used for multi-site brain lesion segmentation [18]. Unlearning the scanner bias [9] is an SDA method for MRI harmonization and improves the performance in age prediction from MR images. The introduction of contrastive loss terms [20,30,25] can also be used for domain generalization [19,23,10]. Disentangling the latent space has been done for MRI harmonization [4,8]. Recently, heterogeneous DA [2] was also of interest for lesion segmentation [5].

1.2 Problem Statement

All DA methods mentioned in Section 1.1 have in common that they must learn the main task on the source domain. However, it can happen that the bias present in datasets of various origins disturbs the learning of a specific task. Figure 1 on the left illustrates the problem and the relation of the different datasets on a toy example for the classification task between hexagons and rectangles. Due to the high variability in data, often only a small and specific dataset is at hand to learn the main task: Dataset 1 forms the target domain with only a small number of samples of hexagons and pentagons. Training on this dataset alone yields a low generalization quality of the model. To increase the number of training samples, we add Dataset 2 and Dataset 3. They form the source domain. As

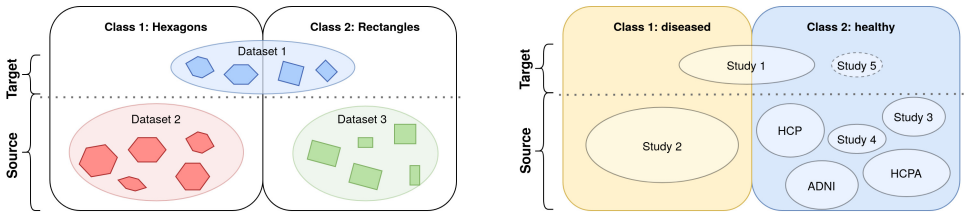


Fig. 1. Quantity charts for the datasets in the source and target domain. The chart on the left illustrates the problem, and the chart on the right shows the real-world application on the MS dataset.

these additional datasets come from different origins, they differ from each other in color. Note that they only provide either rectangles (Dataset 3) or hexagons (Dataset 2). The challenge of such a setup is that during training on the source domain, the color is the dominant feature, and the model learns to distinguish between green and red rather than counting the number of vertices. Classical DA approaches then learn to overcome the domain shift between source and target domain. However, the model will show poor performance on the target domain: The learned features are not helpful, as all hexagons and rectangles are blue in Dataset 1.

This type of problem is highly common in the clinical environment, where different datasets are acquired with different settings, which corresponds to the colors in the toy example. In this project, the main task is to distinguish between multiple sclerosis (MS) patients and healthy controls. The quantity chart in Figure 1 on the right visualizes the different allocations of the MS dataset. Only the small in-house Study 1 provides images of both MS patients and healthy subjects acquired with the same settings. To get more data, we collect images from other in-house studies. As in the hospital mostly data of patients are collected, we add healthy subjects from public datasets, resulting in the presented problem.

In this work, we present a new supervised DA method called *Learn to Ignore (L2I)*, which aims to ignore features related to the scanner bias while focusing on disease-related features for a classification task between healthy and diseased subjects. We exploit the fact that the target domain contains images of subjects of both classes with the same origin, and use this dataset to lead the model's attention to task-specific features. We developed specific constraints on the latent space and introduce two novel loss terms that can be added to any classification network. We evaluate our method on a multi-site MR dataset of MS patients and healthy subjects, compare it to state-of-the-art methods, and perform various ablation studies. The source code is available at `https://gitlab.com/cian.unibas.ch/L2I`.

2 Method

We developed a strategy that aims to ignore features that disturb the learning of a classification task between n classes. The building blocks of our setup are shown in Figure 2. The input image $x_i \in \mathbb{R}^3$ of class $i \in \{1, ..., n\}$ is the input for the encoder network E with parameters θ_{E}, which follows the structure of Inception-ResNet-v1 [26]. However, we replaced the 2D convolutions with 3D convolutions and changed the batch normalization layers to instance normalization layers. The output is the latent vector $f_i = E(x_i) \in \mathbb{R}^m$, where m denotes the dimension of the latent space. This latent vector is normalized to a length of 1 and forms the input for the classification network C with parameters θ_{C}. Finally, we get the classification scores $p_i = C(f_i) = C(E(x_i))$ for class $i \in \{1, ..., n\}$. To make the separation between the classes in the latent space learnable, we introduce additional parameters θ_{O} that learn normalized center points $\mathcal{O} = \{o_1, ..., o_n\} \subset \mathbb{R}^m$.

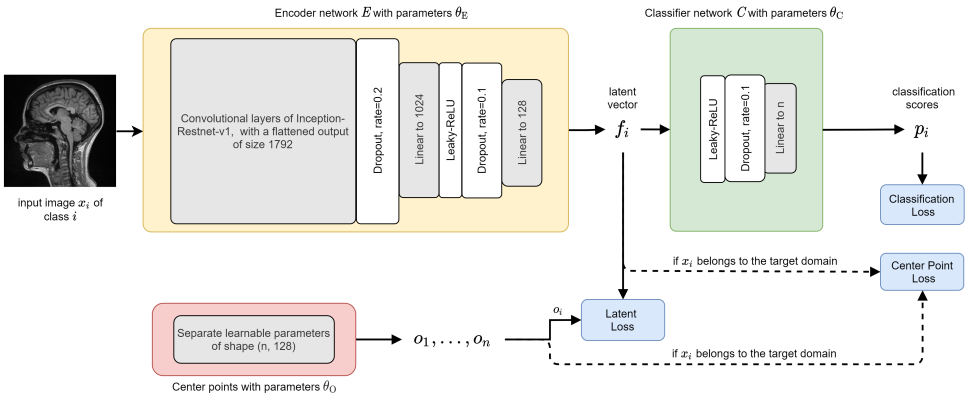


Fig. 2. Architecture of the classification network consisting of an encoder E with parameters θ_E, a fully connected classifier C with parameters θ_C, and separate learnable parameters θ_O. Here, $o_1, ..., o_n$ are learnable center points in the latent space, f_i is a vector in the latent space, and p_i is the classification score for class i.

To suppress the scanner-related features, we embed the latent vectors in the latent space such that latent vectors from the same class are close to each other, and those from different classes are further apart, irrespective of the domain. We exploit the fact that the target domain contains images of all classes of the same origin. The model learns the separation of the embeddings using data of the target domain only. A schematic overview in 2D for the case of $n = 2$ classes is given in Figure 3. We denote the latent vector of an image of the target domain of class i as $f_{i,t}$, where t denotes the affiliation to the target domain. The center points $\mathcal{O}$ are learned considering the latent vectors $f_{i,t}$ only, such that o_i is close to $f_{i,t}$, for $i \in \{1, ..., n\}$, and o_i is far from o_j for $i \neq j$. We force the latent vector f_i of an input image x_i into a hypersphere of radius r centered in o_i. For illustration, we use the toy example of Section 1.2: As all elements of the target domain are blue, the two learnable center points o_1 and o_2 are separated from each other based only on the number of vertices. The color is ignored. All latent vectors of hexagons f_1 of the source domain should lie in a ball around o_1, and all latent vectors of rectangles f_2 should lie in a ball around o_2.

2.1 Loss functions

The overall objective function is given by

$$\mathcal{L}_{\text{total}} = \underbrace{\mathcal{L}_{\text{cls}}}_{\theta_C} + \underbrace{\lambda_{\text{cen}}\mathcal{L}_{\text{cen}}}_{\theta_O,\theta_E} + \underbrace{\lambda_{\text{latent}}\mathcal{L}_{\text{latent}}}_{\theta_E}. \tag{1}$$

It consists of three components: A classification loss $\mathcal{L}_{\text{cls}}$, a center point loss $\mathcal{L}_{\text{cen}}$ for learning $\mathcal{O} = \{o_1, ..., o_n\}$, and a loss $\mathcal{L}_{\text{latent}}$ on the latent space. Those components are weighted with the hyperparameters $\lambda_{\text{latent}}, \lambda_{\text{cen}} \in \mathbb{R}$. The parameters

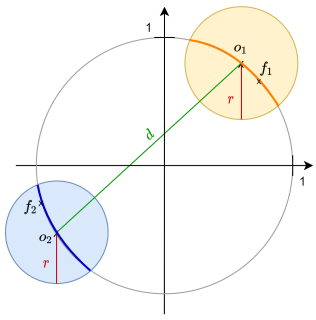


Fig. 3. The diagram shows a 2D sketch of the proposed latent space for $n = 2$ classes, with two learnable center points o_1 and o_2. The latent vectors are normalized and lie on the unit hypersphere. Latent vectors f_i of images of class i should lie within the circle around o_i, on the orange line or blue line respectively.

$\theta_{\mathrm{E}}, \theta_{\mathrm{C}}$ and θ_{O} indicate which parameters of the network are updated with which components of the loss term. While the classification loss $\mathcal{L}_{\mathrm{cls}}$ is separate and only responsible for the final score, it is the center point loss $\mathcal{L}_{\mathrm{cen}}$ and the latent loss $\lambda_{\mathrm{latent}}$ that iteratively adapt the feature space to be scanner-invariant. With this total loss objective, any classification network can be extended by our method.

Classification Loss The classification loss $\mathcal{L}_{\mathrm{cls},\theta_{\mathrm{C}}}(f_i)$ is defined by the cross-entropy loss. The gradient is only calculated with respect to θ_{C}, as we do not want to disturb the parameters θ_{E} with the scanner bias.

Center Point Loss To determine the center points, we designed a novel loss function defined by the distance from a latent vector $f_{i,t}$ of the target domain to its corresponding center point o_i. We define a radius $r > 0$ and force the latent vectors of the target domain $f_{i,t}$ to be within a hypersphere of radius r centered in o_i. Moreover, o_i and o_j should be far enough from each other for $i \neq j$. As o_i is normalized to a length of one, the maximal possible distance between o_i and o_j equals 2. We add a loss term forcing the distance between o_i and o_j to be larger than a distance d. The choice of $d < 2$ and $r > 0$ with $d > 2r$ is closely related to the choice of a margin in conventional contrastive loss terms [25,14]. The network is not penalized for not forcing o_i in the perfect position, but only to an acceptable region, such that the hyperspheres do not overlap. Then, the center point loss used to update the parameters θ_{O} and θ_{E} is given by

$$\mathcal{L}_{\mathrm{cen},\theta_{\mathrm{O}},\theta_{\mathrm{E}}}(f_{i,t}, \mathcal{O}) = \max(\|f_{i,t} - o_i\|_2 - r, 0)^2 + \sum_{k \neq i} \frac{1}{2} \max(d - \|o_k - o_i\|_2, 0)^2. \quad (2)$$

Latent Loss We define the loss on the latent space, similar to the *Center Loss* [30], by the distance from f_i to its corresponding center point o_i

$$\mathcal{L}_{\text{latent},\theta_{\text{E}}}(f_i, \mathcal{O}) = \max(\|f_i - o_i\|_2 - r, 0)^2. \tag{3}$$

With this loss, all latent vectors f_i of the training set of class i are forced to be within a hypersphere of radius r around the center point o_i. This loss is used to update the parameters θ_{E} of the encoder. By choosing $r > 0$, the network is given some leeway to force f_i to an acceptable region around o_i, denoted by the orange and blue lines in Figure 3.

3 Experiments

For the MS dataset, we collected T1-weighted images acquired with 3T MR scanners with the MPRAGE sequence from five different in-house studies. For data privacy concerns, this patient data is not publicly available. Written informed consent was obtained from all subjects enrolled in the studies. All data were coded (i.e. pseudo-anonymized) at the time of the enrollment of the patients. To increase the number of healthy controls, we also randomly picked MPRAGE images from the Alzheimer's Disease Neuroimaging Initiative[5] (ADNI) dataset, the Young Adult Human Connectome Project (HCP) [29] and the Human Connectome Project - Aging (HCPA) [16]. More details of the different studies are given in Table 1 of the supplementary material, including the split into training, validation, and test set. An example of the scanner bias effect for two healthy control groups of the ADNI and HCP dataset can be found in Section 3 of the supplementary material.
All images were preprocessed using the same pipeline consisting of skull-stripping with HD-BET [17], N4 biasfield correction [27], resampling to a voxel size of 1 mm×1 mm×1 mm, cutting the top and lowest two percentiles of the pixel intensities, and finally an affine registration to the MNI 152 standard space[6]. All images were cropped to a size of (124, 120, 172). The dimension of the latent space is $m = 128$. This results in a total number of parameters of $36,431,842$. We use the Adam optimizer [21] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a weight decay of $5 \cdot 10^{-5}$. The learning rate for the parameters θ_{O} is $lr_{\text{O}} = 10^{-4}$, and the learning rate for the parameters θ_{C} and θ_{E} is $lr_{\text{E,C}} = 5 \cdot 10^{-5}$. We manually choose the hyperparameters $\lambda_{\text{latent}} = 1$, $\lambda_{\text{cen}} = 100$, $d = 1.9$, and $r = 0.1$.
An early stopping criterion, with a patience value of 20, based on the validation loss on the target domain, is used. For data sampling in the training set, we use the scheme presented in Algorithm 1 in Section 2 of the supplementary material.

[5] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative(ADNI) database (adni.loni.usc.edu). The investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report.

[6] Copyright (C) 1993-2009 Louis Collins, McConnell Brain Imaging Centre, Montreal Neurological Institute, McGill University.

Data augmentation includes rotation, gamma correction, flipping, scaling, and cropping. The training was performed on an NVIDIA Quadro RTX 6000 GPU, and took about 8 hours on the MS dataset. As software framework, we used Pytorch 1.5.0.

4 Results and Discussion

To measure the classification performance, we calculate the classification accuracy, the Cohen's kappa score [6], and the area under the receiver operating characteristic curve (AUROC) [15]. We compare our approach against the methods listed below. Implementation details and the source code of all comparing methods can be found at `https://gitlab.com/cian.unibas.ch/L2I`.

- Vanilla classifier (*Vanilla*): Same architecture as *L2I*, but only $\mathcal{L}_{\text{cls}}$ is taken to update the parameters of both the encoder E and the classifier C.
- Class-aware Sampling (*Class-aware*): We train the *Vanilla* classifier with class-aware sampling. In every batch each class and domain is represented.
- Weighted Loss (*Weighted*): We train the *Vanilla* classifier, but the loss function $\mathcal{L}_{\text{cls}}$ is weighted to compensate for class and domain imbalances.
- Domain-Adversarial Neural Network (*DANN*) [12]: The classifier learns both to distinguish between the domains and between the classes. It includes a gradient reversal for the domain classification.
- Unlearning Scanner Bias (*Unlearning*) [9]: A confusion loss aims to maximally confuse a domain predictor, such that only features relevant for the main task are extracted.
- Supervised Contrastive Learning (*Contrastive*) [20]: The latent vectors are pushed into clusters far apart from each other, with a sampling scheme and a contrastive loss term allowing for multiple positives and negatives.
- Contrastive Adaptation Network (*CAN*) [19]: This state-of-the-art DA method combines the maximum mean discrepancy of the latent vectors as loss objective, class-aware sampling, and clustering.
- Fixed Center Points (*Fixed*): We train *L2I*, but instead of learning the centerpoints o_1 and o_2 using the target domain, we fix the center points at $o_i = \frac{v_i}{\|v_i\|_2}$ for $i \in \{1, 2\}$, with $v_1 = (1, ..., 1)$ and $v_2 = (-1, ..., -1)$.
- No margin (*No-margin*): We train *L2I* with $d = 2$ and $r = 0$, such that no margin is chosen in the contrastive loss term in Equations 2 and 3.

We report the mean and standard deviation of the scores on the test set for 10 runs. For each run the dataset is randomly divided into training, validation, and test set. In the first three lines of Table 1, the scores are shown when *Vanilla*, *Class-aware*, and *Weighted* are trained only on the target domain. The very poor performance is due to overfitting on such a small dataset. Therefore, the target domain needs to be supplemented with other datasets. In the remaining lines of Table 1, we summarize the classification results for all methods when trained on the source and the target domain. Our method *L2I* strongly outperforms all

Table 1. Mean [standard deviation] of the scores on the test set for 10 runs.

set		Target Domain			Source Domain		
		accuracy	**kappa**	**AUROC**	**accuracy**	**kappa**	**AUROC**
target	Vanilla	50.0 [0.0]	0.0 [0.0]	59.5 [13.0]			
	· Class-aware	65.0 [5.0]	29.3 [9.5]	68.6 [12.4]			
	· Weighted	50.3 [1.1]	0.0 [0.0]	73.3 [13.5]			
target and source	Vanilla	69.0 [6.1]	38.0 [12.1]	79.8 [7.3]	90.3 [5.7]	80.7 [11.3]	95.0 [5.9]
	· Class-aware	71.0 [8.3]	42.0 [16.6]	79.3 [13.3]	90.7[4.2]	81.3 [8.3]	95.3 [3.9]
	· Weighted	71.3 [6.3]	42.7 [12.7]	81.2 [7.8]	92.2 [3.1]	84.6 [6.3]	95.1 [3.4]
	DANN	67.0 [5.7]	34.0 [11.5]	74.7 [8.4]	93.1 [2.8]	86.3 [5.5]	98.7 [0.7]
	Unlearning	70.7 [6.6]	41.3 [13.3]	81.7 [9.2]	85.5 [3.2]	71.0 [6.5]	90.5 [3.3]
	Contrastive	76.3 [6.7]	52.6 [13.5]	86.7 [9.5]	**94.5 [2.4]**	**89.0 [4.7]**	**98.9 [0.7]**
	CAN	75.3 [6.9]	50.7 [13.8]	83.8 [8.8]	92.8 [2.5]	85.7 [5.0]	96.7 [1.8]
	L2I [Ours]	**89.0 [3.9]**	**78.0 [7.7]**	**89.7 [7.6]**	92.0 [2.5]	84.0 [4.9]	91.7 [4.9]
	·Fixed	71.7[7.2]	43.3 [14.5]	76.5 [11.4]	90.5 [3.9]	81.0 [7.9]	90.2 [5.8]
	·No-Margin	82.0 [3.9]	64.0 [7.8]	75.3 9.1]	91.7 [4.8]	83.3 [9.7]	84.9 [3.9]

other methods on the target domain. Although we favor the target domain during training, we see that *L2I* still has a good performance on the source domain. Therefore, we claim that the model learned to distinguish the classes based on disease-related features that are present in both domains, rather than based on scanner-related features. The benefit of learning o_1 and o_2 by taking only the target domain into account can be seen when comparing our method against *Fixed*. Moreover, by comparing *L2I* to *No-margin*, we can see that choosing $d < 2$ and $r > 0$ brings an advantage. All methods perform well on the source domain, where scanner-related features can be taken into account for classification. However, on the target domain, where only disease-related features can be used, the *Vanilla*, *Class-aware* and *Weighted* methods show a poor performance. A visualization of the comparison between the *Vanilla* classifier and our method *L2I* can be found in the t-SNE plots in Section 4 of the supplementary material. *DANN* and *Contrastive*, as well as the state-of-the-art methods *CAN* and *Unlearning* fail to show the performance they achieve in classical DA tasks. Although *CAN* is an unsupervised method, we think that the comparison to our supervised method is fair, since *CAN* works very well on classical DA problems.

5 Conclusion

We presented a method that can ignore image features that are induced by different MR scanners. We designed specific constraints on the latent space and define two novel loss terms, which can be added to any classification network. The novelty lies in learning the center points in the latent space using images of the monocentric target domain only. Consequently, the separation of the latent

space is learned based on task-specific features, also in cases where the main task cannot be learned from the source domain alone. Our problem therefore differs substantially from classical DA or contrastive learning problems. We apply our method *L2I* on a classification task between multiple sclerosis patients and healthy controls on a multi-site MR dataset. Due to the scanner bias in the images, a vanilla classification network and its variations, as well as classical DA and contrastive learning methods, show a weak performance. *L2I* strongly outperforms state-of-the-art methods on the target domain, without loss of performance on the source domain, improving the generalization quality of the model. Medical images acquired with different scanners are a common scenario in long-term or multi-center studies. Our method shows a major improvement for this scenario compared to state-of-the-art methods. We plan to investigate how other tasks like image segmentation will improve by integrating our approach.

References

1. Ackaouy, A., Courty, N., Vallée, E., Commowick, O., Barillot, C., Galassi, F.: Unsupervised domain adaptation with optimal transport in multi-site segmentation of multiple sclerosis lesions from mri data. Frontiers in computational neuroscience **14**, 19 (2020)
2. Alipour, N., Tahmoresnezhad, J.: Heterogeneous domain adaptation with statistical distribution alignment and progressive pseudo label selection. Applied Intelligence pp. 1–18 (2021)
3. Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.W.: A theory of learning from different domains. Machine learning **79**(1), 151–175 (2010)
4. Chartsias, A., Joyce, T., Papanastasiou, G., Semple, S., Williams, M., Newby, D.E., Dharmakumar, R., Tsaftaris, S.A.: Disentangled representation learning in cardiac image analysis. Medical image analysis **58**, 101535 (2019)
5. Chiou, E., Giganti, F., Punwani, S., Kokkinos, I., Panagiotaki, E.: Unsupervised domain adaptation with semantic consistency across heterogeneous modalities for mri prostate lesion segmentation. In: Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health, pp. 90–100. Springer (2021)
6. Cohen, J.: A coefficient of agreement for nominal scales. Educational and Psychological Measurement **20**(1), 37–46 (1960)
7. Dewey, B.E., Zhao, C., Reinhold, J.C., Carass, A., Fitzgerald, K.C., Sotirchos, E.S., Saidha, S., Oh, J., Pham, D.L., Calabresi, P.A., et al.: Deepharmony: a deep learning approach to contrast harmonization across scanner changes. Magnetic resonance imaging **64**, 160–170 (2019)
8. Dewey, B.E., Zuo, L., Carass, A., He, Y., Liu, Y., Mowry, E.M., Newsome, S., Oh, J., Calabresi, P.A., Prince, J.L.: A disentangled latent space for cross-site MRI harmonization. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 720–729. Springer (2020)
9. Dinsdale, N.K., Jenkinson, M., Namburete, A.I.: Unlearning scanner bias for MRI harmonisation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 369–378. Springer (2020)
10. Dou, Q., Coelho de Castro, D., Kamnitsas, K., Glocker, B.: Domain generalization via model-agnostic learning of semantic features. Advances in Neural Information Processing Systems **32** (2019)

11. Eshaghzadeh Torbati, M., Minhas, D.S., Ahmad, G., O'Connor, E.E., Muschelli, J., Laymon, C.M., Yang, Z., Cohen, A.D., Aizenstein, H.J., Klunk, W.E., Christian, B.T., Hwang, S.J., Crainiceanu, C.M., Tudorascu, D.L.: A multi-scanner neuroimaging data harmonization using ravel and combat. NeuroImage **245** (2021)
12. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. J. Mach. Learn. Res. **17**(1), 2096–2030 (Jan 2016)
13. Guan, H., Liu, M.: Domain adaptation for medical image analysis: a survey. IEEE Transactions on Biomedical Engineering (2021)
14. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. vol. 2, pp. 1735–1742 (2006)
15. Hanley, J.A., McNeil, B.J.: The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology **143**(1), 29–36 (1982)
16. Harms, M.P., Somerville, L.H., Ances, B.M., Andersson, J., Barch, D.M., Bastiani, M., Bookheimer, S.Y., Brown, T.B., Buckner, R.L., Burgess, G.C., et al.: Extending the Human Connectome Project across ages: Imaging protocols for the Lifespan Development and Aging projects. NeuroImage **183**, 972–984 (2018)
17. Isensee, F., Schell, M., Pflueger, I., Brugnara, G., Bonekamp, D., Neuberger, U., Wick, A., Schlemmer, H.P., Heiland, S., Wick, W., et al.: Automated brain extraction of multisequence MRI using artificial neural networks. Human brain mapping **40**(17), 4952–4964 (2019)
18. Kamnitsas, K., Baumgartner, C., Ledig, C., Newcombe, V., Simpson, J., Kane, A., Menon, D., Nori, A., Criminisi, A., Rueckert, D., Glocker, B.: Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In: International conference on information processing in medical imaging. pp. 597–609. Springer (2017)
19. Kang, G., Jiang, L., Yang, Y., Hauptmann, A.G.: Contrastive adaptation network for unsupervised domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4893–4902 (2019)
20. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. arXiv preprint arXiv:2004.11362 (2020)
21. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
22. Mårtensson, G., Ferreira, D., Granberg, T., Cavallin, L., Oppedal, K., Padovani, A., Rektorova, I., Bonanni, L., Pardini, M., Kramberger, M.G., et al.: The reliability of a deep learning model in clinical out-of-distribution mri data: a multicohort study. Medical Image Analysis **66**, 101714 (2020)
23. Motiian, S., Piccirilli, M., Adjeroh, D.A., Doretto, G.: Unified deep supervised domain adaptation and generalization. In: Proceedings of the IEEE international conference on computer vision. pp. 5715–5725 (2017)
24. Sankaranarayanan, S., Balaji, Y., Castillo, C.D., Chellappa, R.: Generate to adapt: Aligning domains using generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8503–8512 (2018)
25. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 815–823 (2015)

26. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 31 (2017)
27. Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C.: N4ITK: improved N3 bias correction. IEEE transactions on medical imaging **29**(6), 1310–1320 (2010)
28. Valverde, S., Salem, M., Cabezas, M., Pareto, D., Vilanova, J.C., Ramió-Torrentà, L., Rovira, À., Salvi, J., Oliver, A., Lladó, X.: One-shot domain adaptation in multiple sclerosis lesion segmentation using convolutional neural networks. NeuroImage: Clinical **21**, 101638 (2019)
29. Van Essen, D.C., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T.E., Bucholz, R., Chang, A., Chen, L., Corbetta, M., Curtiss, S.W., et al.: The Human Connectome Project: a data acquisition perspective. Neuroimage **62**(4), 2222–2231 (2012)
30. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: European conference on computer vision. pp. 499–515. Springer (2016)

Supplementary Material

1 Study details of the MS dataset

Table 1. Overview of the studies in the MS dataset. All images are T1-weighted MPRAGE images acquired with a 3T Siemens scanner. The dataset is split proportionally into training, validation and test set. Since there are only few images of healthy subjects in Study 1, we add 9 images of healthy subjects of Study 5. Those images were also acquired on the same scanner, but they differ in the acceleration factor from images of Study 1. To make sure that this slight difference does not disturb the training or provide biased results during test time, we use those images only for validation.

		Training		**Validation**		**Testing**		**Scanner**	**Age**	**Gender[m/f]**
		healthy	MS	healthy	MS	healthy	MS			
Target	Study 1	16	55		9	15	15	Trio Trim	19-62	34 / 58
	Study 5			9				Trio Trim	26-55	4 / 5
Source	Study 2		368		20		30	Skyra-fit, Skyra	20-76	137 / 279
	ADNI	113		6		9		Trio Trim	56-78	54 / 74
	HCP	89		4		7		Connectome	22-35	49 / 51
	HCPA	140		7		10		Prisma	36-88	66 / 91
	Study 3	40		2		3		Skyra	27-78	21 / 31
	Study 4	22		1		1		Skyra	21-47	12 / 12

2 Sampling algorithm on the MS dataset

Algorithm 1 Sampling scheme

repeat

- Sample a random batch from the whole training set (target and source domain) of size 10. Use this batch to calculate $\mathcal{L}_{\text{cls}}$ and $\mathcal{L}_{\text{latent}}$.
- Sample one image from the target domain of each class and calculate $\mathcal{L}_{\text{center}}$.
- Update network parameters with $\mathcal{L}_{\text{total}}$.

until early stopping criterion on validation loss is reached.

3 Histogram showing the scanner bias effect

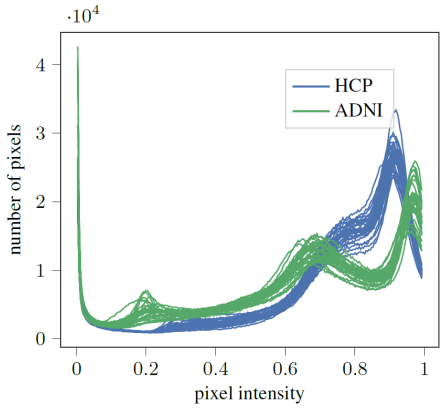


Fig. 1. Distribution of the pixel intensities for preprocessed and normalized MPRAGE MR images of healthy subjects of the ADNI and the HCP dataset. It can clearly be seen that the distribution of the pixel intensities is still different for each dataset after preprocessing. This difference originates from variations in scanner hardware, software version or scanning protocols.

4 T-SNE plots

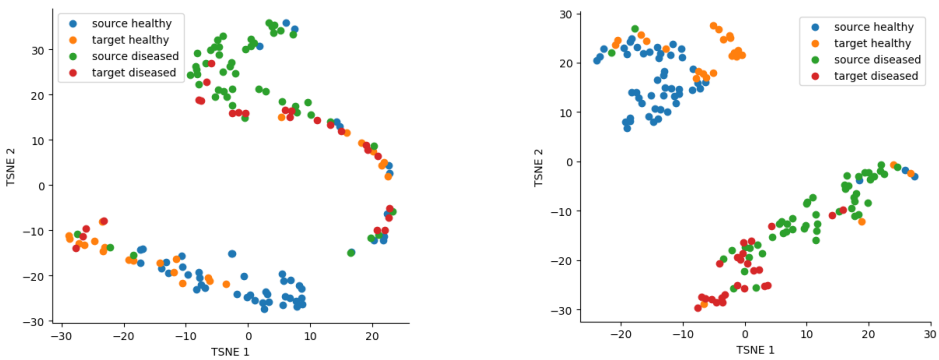


Fig. 2. On the left, we present the t-SNE plot of the latent vectors using the *Vanilla* classifier. On the target domain (orange and red datapoints), the distinction between healthy and diseased subjects cannot clearly be seen. On the right, we show the t-SNE plot of the latent vectors using *L2I*. The separation between healthy and diseased subjects can be seen, regardless of the affiliation to the source or target domain.

Chapter 10

Discussion and Conclusion

The aim of this work was the development of methods for the automatic detection and segmentation of pathological regions in medical images. Limited availability of data and corresponding labels imposes various challenges to state-of-the-art deep learning approaches and has a direct impact on the degree of automatization that can be achieved. We explored various scenarios depending on the amount of data and labels accessible and proposed methods to cope with the given limitations. An overview of the different scenarios and the developed methods is shown in Figure 1.3.

Deep learning models have shown excellent performance for image segmentation when trained in a fully supervised setting. However, manual segmentations, which are required for this type of training, are underlying variations and prone to human bias. Inter-rater variability and unclear anatomical outlines, e.g., blurry outlines of brain tumors, can lead to inaccurate ground truth segmentations. Moreover, there is uncertainty in the model parameter themselves. For clinical applications, it is essential to identify the regions with high certainty in the predicted segmentation. In order to highlight these regions, we presented a novel segmentation approach based on diffusion models. We evaluated our novel method on the BRATS2020 challenge for a binary tumor segmentation task. Compared to classical fully supervised segmentation methods, where the segmentation is computed in one step, the generation of the segmentation mask was performed by a Markov process described in Section 3.3.2. Applying this stochastic generation process multiple times for each input image, this implicitly provided an ensemble of segmentation masks. Already an ensemble of five segmentation masks resulted in segmentation scores close to state-of-the-art. The variance map of this ensemble provided pixel-wise uncertainty maps, for example, supporting the experts in finding a safety margin for tumor resection. We visually compared these uncertainty maps to the aleatoric and epistemic uncertainty maps using Bayesian neural networks, showing that similar regions are highlighted. In contrast to these methods, however, our approach does not rely on adding noise to the input image or masking model parameters. The uncertainty originates from the stochasticity in the generation process itself. A quantitative analysis of the resulting uncertainty maps is yet to be explored.

By training on pixel-wise manual labels, fully supervised segmentation methods learn to imitate human performance. Moreover, such pixel-wise ground truth labels are often unavailable. Therefore, we focused on weakly supervised anomaly detection based on image-to-image

translation to obtain pixel-wise anomaly maps. Given two datasets, one containing patients and one containing healthy controls, we translated images showing a pathology to images showing no pathology. The anomaly map was defined by the difference between the two. In PathoGAN [5], Fixed-Point-GAN [168] and VAGAN [14], GANs were adapted to perform this type of image-to-image translations task. However, a challenge arises when the pathology is defined by a structural deformation of already existing structures rather than lesions. In this case, the generation of an inpainting or an additive map, as proposed by the previous approaches, is insufficient to capture the difference between the data distributions of the healthy and the diseased subjects. We defined a new method called *DeScarGAN*, where a new image is generated without the constraints of an inpainting or additive approach. This provided more freedom to the generation process and opened the possibility of dealing with structural deformations. The major challenge was that the GAN must preserve the identity in regions showing healthy tissue or background. This ensures that the output of the model does not show *any* subject but the healthy reconstruction of the input image, which is of great importance for the anomaly map to be accurate. In *DeScarGAN*, we proposed additional identity-preserving loss terms and an architecture with skip connections to ensure this identity-preserving mechanism. For evaluation, we designed a specific synthetic dataset with pixel-wise ground truth, where the anomalous images show in shrinkage of specific image regions. For the medical applications, we visually compared the results on X-ray images of the lungs, where the patients suffer from pleural effusions. The flexibility of generating a whole new image, combined with the proposed detail-preserving mechanisms, enabled us to outperform the comparing methods. However, the extension to 3D, which is essential for processing MR or CT images, did not yield good results. Memory requirements, long training times and the unstable GAN training are the major reasons that prevented us from moving from 2D to 3D.

Due to the complex and unstable training of such an identity preserving GAN, we replaced the GAN with a denoising diffusion model. A key advantage is that no adversarial training is involved; only an MSE loss is taken to update a model with a U-Net architecture. This renders the training process straightforward. The resulting generated images are of excellent quality: A short time after DDPMs were introduced, it was shown that diffusion models can beat GANs at what they do best, i.e., the generation of fake images [42]. We took advantage of this promising new method and adapted it to an image-to-image translation task. The deterministic iterative noising and denoising process of denoising diffusion implicit models encodes input images in noise, thereby guaranteeing the detail-preserving mechanism during image-to-image translation. Therefore, no additional constraints need to be made during training, which makes it less prone to error. To achieve our goal of translating an input image of a patient to its healthy reconstruction, an external classifier was trained to distinguish between healthy and diseased subjects. We included classifier guidance in the denoising process, initially proposed to guide the generation of synthetic images towards the desired class. To the best of our knowledge, we are the first to use this process for detail-preserving image-to-image translation. Thereby, the images are only changed in regions that contribute to the abnormality of the input image. Consequently, the resulting anomaly maps are very clear compared to other anomaly detection methods. We want to highlight that this happens only during evaluation, and the stable training process of diffusion models remains unchanged. A drawback is the iterative generation process,

which takes longer than the GAN-based approach. This may represent a significant challenge when applying diffusion models to 3D data.
This simple combination of a diffusion model with an external, task-specific network opened the door for many other applications. Having built the diffusion model for anomaly detection based on a binary classification network, we extended this idea. We included the gradient of a regression model and a segmentation model to generate images matching the desired output characteristics. This resulted in a very flexible framework for image-to-image translation, which can be used for the simulation of tumor growth, atrophy, or the generation of data for evaluating anomaly detection methods. One significant advantage is that the diffusion model and the external network are trained separately. Consequently, the same diffusion model can be used for various image-to-image translation tasks on the same dataset. Other image-to-image translation methods using the DDIM encoding and decoding scheme do not rely on such an external network. They need to change the training scheme of the diffusion models to perform image-to-image translation, and consequently, they need to be retrained for each specific task [91, 136].
Finally, all the methods mentioned above are only as good as the data they are trained on. MR images originating from different scanners are biased towards the acquisition settings. This may result in a challenging combination of datasets of multiple sites, limiting the performance of a deep learning model. We presented a novel domain adaptation method *L2I* that learns to ignore the scanner-related features in the images while learning features relevant to a classification task. We focused on a data setup where only a small dataset provided images of all classes. Due to this challenging setup, state-of-the-art domain adaptation methods for MR harmonization failed to show the performance they have on classical domain adaptation tasks. Our method showed a robust performance across all scanners. We argue that mechanisms such as the proposed one must be included in more advanced tasks if we want reliable and generalizable deep learning methods in the clinic.

Future work

With the current encouraging results using diffusion models, a wide field of clinical applications opened. Our preliminary work can be extended to more complex image-to-image translation tasks. Using a longitudinal dataset, a process such as tumor growth could be simulated using an adapted regression approach. In this scope, brain atrophy could also be modeled during the aging process. Another goal is to simultaneously change multiple attributes of one image, e.g., if a patient suffers from multiple diseases, using image-to-image translation. This could be achieved by extending the binary classification used in the weakly supervised anomaly detection approaches to a multi-class classification problem, or by applying the gradients of multiple external networks.
Regarding the fully supervised segmentation approach, we made a simplification to a binary classifiation problem: The tumor class comprised the GD-enhancing tumor, the peritumoral edema, and the necrotic and non-enhancing tumor core. The next step would be a multi-class segmentation for a more detailed analysis of those subclasses. For this, experiments with multi-

channel outputs or non-binary ground truth segmentations need to be considered. Furthermore, a quantitative analysis of the uncertainty maps is required for comparison against other pixel-wise uncertainty estimation methods. For example, it would be interesting to compute our method's calibration error.
Since the proposed approaches are implemented in 2D, an extension to 3D will be necessary for many medical problems. This requires the development of efficient and lean methods that fit the memory restrictions of the GPUs. Moreover, especially regarding diffusion models, there is much potential to reduce the evaluation time.
Since data is often limited in many real-world problems, a multi-site setting such as the one presented in Chapter 9 is very likely to occur. Also, in setups like federated learning, some harmonization mechanism needs to be included in the pipeline to ensure the generalizability of the model to different sites. Any classification network can be adapted with our domain adaptation approach *L2I*. The weakly supervised anomaly detection methods presented in Chapters 6 and 7 both include a binary classification network. It is crucial that this classifier is robust and performs well across scanners. The next step is to combine our anomaly detection approaches with the scanner-invariant classifier to get generalizable and stable methods. It is of interest to investigate how our domain adaptation approach *L2I* integrates into more advanced tasks such as segmentation or anomaly detection.
In this work, we relied either on fully labeled data or on a weakly labeled dataset of patients as well as healthy controls. With even less information available, the obvious extension of this thesis would be to deal with data without labels. For example, if only healthy controls were available, unsupervised or self-supervised anomaly detection methods would need to be explored.

Conclusion

The fast progress of deep learning opened many possibilities for the automatization of medical image analysis. In this thesis, we had to cope with various settings related to the availability of data or labels in medical applications. While the overall goal was to find pathological regions in images of patients automatically, we explored different scenarios relating to the amount of data and type of labels available.
Our main contribution is the adaptation of generative models to different medical applications. We extended denoising diffusion models for the segmentation of brain tumors. Our novel method enabled us to compute uncertainty maps of the segmentation. Furthermore, we could show that generative adversarial nets as well as denoising diffusion models can be used for weakly supervised anomaly detection in medical images. Our methods have shown great performance even in cases where the pathology shows in deformation of already existing structures. Based on denoising diffusion implicit models, we proposed a novel and flexible framework for detail-preserving image-to-image translation for medical image analysis. It showed exceptional results for various tasks, e.g., tumor growth. Finally, we presented a domain adaptation method for robust MR analysis in a multi-site setting.
All presented methods perform well for their given task, each representing a building block toward generalizable and applicable deep learning methods. While all our proposed approaches show promising results, it will be an inevitable step to put the building blocks together to

improve robustness and interpretability to establish reliable deep learning methods in the clinic. We want to highlight the role of diffusion models in this thesis, which offer an interesting alternative to generative adversarial networks for image generation and opened a new field of research.

Bibliography

[1] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon. Ganomaly: Semi-supervised anomaly detection via adversarial training. In *Asian conference on computer vision*, pages 622–637. Springer, 2018.

[2] S. Albelwi. Survey on self-supervised learning: Auxiliary pretext tasks and contrastive learning methods in imaging. *Entropy*, 24(4):551, 2022.

[3] E. Alpaydin. *Introduction to Machine Learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 3 edition, 2014.

[4] R. Anantharaman, M. Velazquez, and Y. Lee. Utilizing mask r-cnn for detection and segmentation of oral diseases. In *2018 IEEE international conference on bioinformatics and biomedicine (BIBM)*, pages 2197–2204. IEEE, 2018.

[5] S. Andermatt, A. Horváth, S. Pezold, and P. Cattin. Pathology segmentation using distributional differences to images of healthy origin. In *International MICCAI Brainlesion Workshop*, pages 228–238. Springer, 2018.

[6] S. Andermatt, S. Pezold, and P. C. Cattin. Automated segmentation of multiple sclerosis lesions using multi-dimensional gated recurrent units. In *International MICCAI Brainlesion Workshop*, pages 31–42. Springer, 2017.

[7] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

[8] N. T. Arun, N. Gaw, P. Singh, K. Chang, K. V. Hoebel, J. Patel, M. Gidwani, and J. Kalpathy-Cramer. Assessing the validity of saliency maps for abnormality localization in medical imaging. *arXiv preprint arXiv:2006.00063*, 2020.

[9] R. Ashikaga, Y. Araki, and O. Ishida. MRI of head injury using flair. *Neuroradiology*, 39(4):239–242, 1997.

[10] J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[11] R. Bachmann, R. Reilmann, W. Schwindt, H. Kugel, W. Heindel, and S. Krämer. Flair imaging for multiple sclerosis: a comparative mr study at 1.5 and 3.0 tesla. *European radiology*, 16(4):915–921, 2006.

[12] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani, and C. Davatzikos. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Scientific data*, 4(1):1–13, 2017.

[13] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, R. T. Shinohara, C. Berger, S. M. Ha, M. Rozycki, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. *arXiv preprint arXiv:1811.02629*, 2018.

[14] C. F. Baumgartner, L. M. Koch, K. C. Tezcan, J. X. Ang, and E. Konukoglu. Visual feature attribution using wasserstein GANs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8309–8319, 2018.

[15] C. Baur, R. Graf, B. Wiestler, S. Albarqouni, and N. Navab. Steganomaly: inhibiting cyclegan steganography for unsupervised anomaly detection in brain MRI. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 718–727. Springer, 2020.

[16] F. Bieder, J. Wolleb, R. Sandkühler, and P. C. Cattin. Position regression for unsupervised anomaly detection. In *International Conference on Medical Imaging with Deep Learning*, pages 160–172. PMLR, 2022.

[17] B. Billot, D. Greve, K. Van Leemput, B. Fischl, J. E. Iglesias, and A. V. Dalca. A learning strategy for contrast-agnostic MRI segmentation. *arXiv preprint arXiv:2003.01995*, 2020.

[18] S. Bond-Taylor, A. Leach, Y. Long, and C. G. Willcocks. Deep generative modelling: A comparative review of VAEs, GANs, normalizing flows, energy-based and autoregressive models. *arXiv preprint arXiv:2103.04922*, 2021.

[19] M. Brant-Zawadzki, G. D. Gillan, and W. R. Nitz. Mp rage: a three-dimensional, t1-weighted, gradient-echo sequence–initial experience in the brain. *Radiology*, 182(3):769–775, 1992.

[20] A. Brock, J. Donahue, and K. Simonyan. Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[21] W. J. Brownlee, T. A. Hardy, F. Fazekas, and D. H. Miller. Diagnosis of multiple sclerosis: progress and challenges. *The Lancet*, 389(10076):1336–1346, 2017.

[22] S. Cai, Y. Shu, G. Chen, B. C. Ooi, W. Wang, and M. Zhang. Effective and efficient dropout for deep convolutional neural networks. *arXiv preprint arXiv:1904.03392*, 2019.

[23] K. S. Caldemeyer and K. A. Buckwalter. The basic principles of computed tomography and magnetic resonance imaging. *Journal of the American Academy of Dermatology*, 41(5):768–771, 1999.

[24] 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*, 148:77–102, 2017.

[25] J.-R. Chang, M.-S. Wu, W.-H. Yu, C.-C. Chen, C.-K. Yang, Y.-Y. Lin, and C.-Y. Yeh. Stain mix-up: Unsupervised domain generalization for histopathology images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 117–126. Springer, 2021.

[26] C. Charalampidis, A. Youroukou, G. Lazaridis, S. Baka, I. Mpoukovinas, V. Karavasilis, I. Kioumis, G. Pitsiou, A. Papaiwannou, A. Karavergou, et al. Pleura space anatomy. *Journal of thoracic disease*, 7(Suppl 1):S27, 2015.

[27] G. B. Chavhan. *MRI made easy*. JP Medical Ltd, 2013.

[28] C. Chen, Q. Dou, H. Chen, J. Qin, and P.-A. Heng. Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 865–872, 2019.

[29] J. Chen, H. Shao, and C. Hu. Image segmentation based on mathematical morphological operator. *Colorimetry and Image Processing*, 2017.

[30] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[31] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[32] J. Choi, S. Kim, Y. Jeong, Y. Gwon, and S. Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938*, 2021.

[33] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018.

[34] M. J. Christ and R. Parvathi. Segmentation of medical image using clustering and watershed algorithms. *American Journal of Applied Sciences*, 8(12):1349, 2011.

[35] A. Compston and A. Coles. Multiple sclerosis. *The Lancet*, 372(9648):1502–1517, 2008.

[36] H. J. Crayton and H. S. Rossman. Managing the symptoms of multiple sclerosis: A multimodal approach. *Clinical Therapeutics*, 28(4):445–460, 2006.

[37] W. Cui, Y. Liu, Y. Li, M. Guo, Y. Li, X. Li, T. Wang, X. Zeng, and C. Ye. Semi-supervised brain lesion segmentation with an adapted mean teacher model. In *International Conference on Information Processing in Medical Imaging*, pages 554–565. Springer, 2019.

[38] H. P. D'Agostino and M. A. Edens. Physiology, pleural fluid. *StatPearls [Internet]*, 2020.

[39] L. Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[40] H. Deutschmann, N. Hinteregger, U. Wießpeiner, M. Kneihsl, S. Fandler-Höfler, M. Michenthaler, C. Enzinger, E. Hassler, S. Leber, and G. Reishofer. Automated MRI perfusion-diffusion mismatch estimation may be significantly different in individual patients when using different software packages. *European Radiology*, 31(2):658–665, 2021.

[41] B. E. Dewey, C. Zhao, J. C. Reinhold, A. Carass, K. C. Fitzgerald, E. S. Sotirchos, S. Saidha, J. Oh, D. L. Pham, P. A. Calabresi, et al. Deepharmony: A deep learning approach to contrast harmonization across scanner changes. *Magnetic resonance imaging*, 64:160–170, 2019.

[42] P. Dhariwal and A. Nichol. Diffusion models beat GANs on image synthesis. *Advances in Neural Information Processing Systems*, 34, 2021.

[43] F. Di Mattia, P. Galeone, M. De Simoni, and E. Ghelfi. A survey on GANs for anomaly detection. *arXiv preprint arXiv:1906.11632*, 2019.

[44] N. K. Dinsdale, M. Jenkinson, and A. I. Namburete. Deep learning-based unlearning of dataset bias for MRI harmonisation and confound removal. *NeuroImage*, 228:117689, 2021.

[45] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[46] V. Dumoulin and F. Visin. A guide to convolution arithmetic for deep learning. *arXiv preprint arXiv:1603.07285*, 2016.

[47] W. H. Elmasry, H. M. Moftah, N. El-Bendary, and A. E. Hassanien. Graph partitioning based automatic segmentation approach for ct scan liver images. In *2012 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pages 183–186. IEEE, 2012.

[48] M. Eshaghzadeh Torbati, D. S. Minhas, G. Ahmad, E. E. O'Connor, J. Muschelli, C. M. Laymon, Z. Yang, A. D. Cohen, H. J. Aizenstein, W. E. Klunk, B. T. Christian, S. J. Hwang, C. M. Crainiceanu, and D. L. Tudorascu. A multi-scanner neuroimaging data harmonization using ravel and combat. *NeuroImage*, 245:118703, 2021.

[49] A. Farahani, S. Voghoei, K. Rasheed, and H. R. Arabnia. A brief review of domain adaptation. *Advances in Data Science and Information Engineering*, pages 877–894, 2021.

[50] F. Farnia and A. Ozdaglar. Do GANs always have nash equilibria? In *International Conference on Machine Learning*, pages 3029–3039. PMLR, 2020.

[51] T. Fernando, H. Gammulle, S. Denman, S. Sridharan, and C. Fookes. Deep learning for medical anomaly detection–a survey. *arXiv preprint arXiv:2012.02364*, 2020.

[52] M. A. Flower. *Webb's physics of medical imaging*. CRC press, 2012.

[53] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

[54] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, 2016.

[55] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[56] K. Gregor, I. Danihelka, A. Graves, D. Rezende, and D. Wierstra. Draw: A recurrent neural network for image generation. In *International Conference on Machine Learning*, pages 1462–1471. PMLR, 2015.

[57] H. Guan and M. Liu. Domain adaptation for medical image analysis: a survey. *IEEE Transactions on Biomedical Engineering*, 2021.

[58] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of wasserstein GANs. *Advances in neural information processing systems*, 30, 2017.

[59] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006.

[60] E. Hattingen, A. Müller, A. Jurcoane, B. Mädler, P. Ditter, H. Schild, U. Herrlinger, M. Glas, and S. Kebir. Value of quantitative magnetic resonance imaging t1-relaxometry in predicting contrast-enhancement in glioblastoma patients. *Oncotarget*, 8(32):53542, 2017.

[61] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[62] G. Hinton, N. Srivastava, and K. Swersky. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. *Cited on*, 14(8):2, 2012.

[63] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[64] H. Hojjati, T. K. K. Ho, and N. Armanfard. Self-supervised anomaly detection: A survey and outlook. *arXiv preprint arXiv:2205.05173*, 2022.

[65] A. Horváth, C. Tsagkas, S. Andermatt, S. Pezold, K. Parmar, and P. Cattin. Spinal cord gray matter-white matter segmentation on magnetic resonance amira images with md-gru. In *International Workshop and Challenge on Computational Methods and Clinical Applications for Spine Imaging*, pages 3–14. Springer, 2018.

[66] E. Hosseini-Asl, R. Keynton, and A. El-Baz. Alzheimer's disease diagnostics by adaptation of 3d convolutional network. In *2016 IEEE International Conference on Image Processing*, pages 126–130, 2016.

[67] S.-W. Huang, C.-T. Lin, S.-P. Chen, Y.-Y. Wu, P.-H. Hsu, and S.-H. Lai. Auggan: Cross domain adaptation with GAN-based data augmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 718–731, 2018.

[68] S.-W. Huang, C.-T. Lin, S.-P. Chen, Y.-Y. Wu, P.-H. Hsu, and S.-H. Lai. AugGAN: Cross domain adaptation with GAN-based data augmentation. In *Proceedings of the European Conference on Computer Vision*, pages 718–731, 2018.

[69] X. Huang and S. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017.

[70] M. IJzerman-Korevaar, T. J. Snijders, A. de Graeff, S. C. Teunissen, and F. Y. de Vos. Prevalence of symptoms in glioma patients throughout the disease trajectory: a systematic review. *Journal of neuro-oncology*, 140(3):485–496, 2018.

[71] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

[72] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.

[73] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

[74] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[75] B. Jany and T. Welte. Pleural effusion in adults—etiology, diagnosis, and treatment. *Deutsches Ärzteblatt International*, 116(21):377, 2019.

[76] M. Jensen and J. E. Wilhjelm. X-ray imaging: Fundamentals and planar imaging, 2006.

[77] J. Jiang, Y. Hu, N. Tyagi, P. Zhang, A. Rimner, G. Mageras, J. Deasy, and H. Veeraraghavan. *Tumor-Aware, Adversarial Domain Adaptation from CT to MRI for Lung Cancer Segmentation*, volume 11071, pages 777–785. 09 2018.

[78] I. Jovčevska, N. Kočevar, and R. Komel. Glioma and glioblastoma-how much do we (not) know? *Molecular and clinical oncology*, 1(6):935–941, 2013.

[79] A. Jungo and M. Reyes. Assessing reliability and challenges of uncertainty estimations for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 48–56. Springer, 2019.

[80] H. Kalinic. Atlas-based image segmentation: A survey. *Croatian Scientific Bibliography*, pages 1–7, 2009.

[81] K. Kamnitsas, C. Baumgartner, C. Ledig, V. Newcombe, J. Simpson, A. Kane, D. Menon, A. Nori, A. Criminisi, D. Rueckert, et al. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In *International conference on information processing in medical imaging*, pages 597–609. Springer, 2017.

[82] G. Kang, L. Jiang, Y. Yang, and A. G. Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4893–4902, 2019.

[83] O. Kargiotis, A. Paschali, L. Messinis, and P. Papathanasopoulos. Quality of life in multiple sclerosis: effects of current treatment options. *International review of psychiatry*, 22(1):67–82, 2010.

[84] V. S. Karkhanis and J. M. Joshi. Pleural effusion: diagnosis, treatment, and management. *Open access emergency medicine: OAEM*, 4:31, 2012.

[85] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863, 2021.

[86] H. Kasban, M. El-Bendary, and D. Salama. A comparative study of medical imaging techniques. *International Journal of Information Science and Intelligent System*, 4(2):37–58, 2015.

[87] G. Katti, S. A. Ara, and A. Shireen. Magnetic resonance imaging (MRI)–a review. *International journal of dental clinics*, 3(1):65–70, 2011.

[88] B. Kaur, P. Lemaître, R. Mehta, N. M. Sepahvand, D. Precup, D. Arnold, and T. Arbel. Improving pathological structure segmentation via transfer learning across diseases. In *Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data*, pages 90–98. Springer, 2019.

[89] A. Kendall and Y. Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

[90] P. M. Kidd. Multiple sclerosis, an autoimmune inflammatory disease: prospects for its integrative management. *Alternative medicine review*, 6(6):540–566, 2001.

[91] G. Kim, T. Kwon, and J. C. Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2426–2435, 2022.

[92] R. King. Atlas of MS 3rd edition. http://www.msif.org/wp-content/uploads/2014/09/Atlas-of-MS.pdf, 2020. Accessed on 05.04.2022.

[93] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[94] D. P. Kingma and M. Welling. An introduction to variational autoencoders. *arXiv preprint arXiv:1906.02691*, 2019.

[95] R. Kohler. A segmentation system based on thresholding. *Computer Graphics and Image Processing*, 15(4):319–338, 1981.

[96] K. M. M. Koriem. Multiple sclerosis: New insights and trends. *Asian Pacific Journal of Tropical Biomedicine*, 6(5):429–440, 2016.

[97] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[98] Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, and F. Huang. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.

[99] Y. Lei, T. Wang, Y. Liu, K. Higgins, S. Tian, T. Liu, H. Mao, H. Shim, W. J. Curran, H.-K. Shu, et al. MRI-based synthetic CT generation using deep convolutional neural network. In *Medical Imaging 2019: Image Processing*, volume 10949, page 109492T. International Society for Optics and Photonics, 2019.

[100] X. Li, L. Yu, H. Chen, C.-W. Fu, L. Xing, and P.-A. Heng. Transformation-consistent self-ensembling model for semisupervised medical image segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2):523–534, 2020.

[101] R. Light. Diagnostic principles in pleural disease. *European Respiratory Journal*, 10(2):476–481, 1997.

[102] R. W. Light. Pleural effusion. *New England Journal of Medicine*, 346(25):1971–1977, 2002.

[103] X. Liu, D. H. Park, S. Azadi, G. Zhang, A. Chopikyan, Y. Hu, H. Shi, A. Rohrbach, and T. Darrell. More control for free! image synthesis with semantic diffusion guidance. *arXiv preprint arXiv:2112.05744*, 2021.

[104] G. Lloyd-Jones. Basics of x-ray physics. https://www.radiologymasterclass.co.uk/tutorials/physics/x-ray_physics_introduction, 2016. Accessed on 05.04.2022.

[105] 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*, pages 3431–3440, 2015.

[106] D. N. Louis, A. Perry, G. Reifenberger, A. Von Deimling, D. Figarella-Branger, W. K. Cavenee, H. Ohgaki, O. D. Wiestler, P. Kleihues, and D. W. Ellison. The 2016 world health organization classification of tumors of the central nervous system: a summary. *Acta neuropathologica*, 131(6):803–820, 2016.

[107] A. A. Lovelace. *Notes by A.A. L.* Taylor's Scientific Memoirs, 1843.

[108] Y. Lu and P. Xu. Anomaly detection for skin disease images using variational autoencoder. *arXiv preprint arXiv:1807.01349*, 2018.

[109] F. D. Lublin, S. C. Reingold, J. A. Cohen, G. R. Cutter, P. S. Sørensen, A. J. Thompson, J. S. Wolinsky, L. J. Balcer, B. Banwell, F. Barkhof, et al. Defining the clinical course of multiple sclerosis: the 2013 revisions. *Neurology*, 83(3):278–286, 2014.

[110] X. Luo, J. Chen, T. Song, and G. Wang. Semi-supervised medical image segmentation through dual-task consistency. pages 8801–8809, 2021.

[111] Q. Ma, T. Zhang, M. V. Zanetti, H. Shen, T. D. Satterthwaite, D. H. Wolf, R. E. Gur, Y. Fan, D. Hu, G. F. Busatto, et al. Classification of multi-site MR images in the presence of heterogeneity using multi-task learning. *NeuroImage: Clinical*, 19:476–486, 2018.

[112] P. Malhotra, S. Gupta, D. Koundal, A. Zaguia, and W. Enbeyle. Deep neural networks for medical image segmentation. *Journal of Healthcare Engineering*, 2022, 2022.

[113] S. N. Marimont and G. Tarroni. Anomaly detection through latent space restoration using vector quantized variational autoencoders. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1764–1767. IEEE, 2021.

[114] A. J. Maubon, J.-M. Ferru, V. Berger, M. C. Soulage, M. DeGraef, P. Aubas, P. Coupeau, E. Dumont, and J.-P. Rouanet. Effect of field strength on mr images: comparison of the same subject at 0.5, 1.0, and 1.5 t. *Radiographics*, 19(4):1057–1067, 1999.

[115] J. McCarthy. What is artificial intelligence? https://www-formal.stanford.edu/jmc/whatisai.pdf, 2004. Accessed on 25.04.2022.

[116] W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

[117] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014.

[118] M. I. Meyer, E. de la Rosa, N. Pedrosa de Barros, R. Paolella, K. Van Leemput, and D. M. Sima. A contrast augmentation approach to improve multi-scanner generalization in mri. *Frontiers in neuroscience*, page 1048, 2021.

[119] D. H. Miller, D. T. Chard, and O. Ciccarelli. Clinically isolated syndromes. *The Lancet Neurology*, 11(2):157–169, 2012.

[120] A. M. Molinaro, J. W. Taylor, J. K. Wiencke, and M. R. Wrensch. Genetic and molecular epidemiology of adult diffuse glioma. *Nature Reviews Neurology*, 15(7):405–417, 2019.

[121] J. P. Mugler III and J. R. Brookeman. Three-dimensional magnetization-prepared rapid gradient-echo imaging (3d mp rage). *Magnetic resonance in medicine*, 15(1):152–157, 1990.

[122] J. Mukhoti, J. van Amersfoort, P. H. Torr, and Y. Gal. Deep deterministic uncertainty for semantic segmentation. *arXiv preprint arXiv:2111.00079*, 2021.

[123] M. Munir, M. A. Chattha, A. Dengel, and S. Ahmed. A comparative analysis of traditional and deep learning-based anomaly detection methods for streaming data. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 561–566. IEEE, 2019.

[124] S. C. Murthy, I. Okereke, D. P. Mason, and T. W. Rice. A simple solution for complicated pleural effusions. *Journal of Thoracic Oncology*, 1(7):697–700, 2006.

[125] F. Nelson, A. Poonawalla, P. Hou, J. Wolinsky, and P. Narayana. 3d mprage improves classification of cortical lesions in multiple sclerosis. *Multiple Sclerosis Journal*, 14(9):1214–1219, 2008.

[126] A. T. Nguyen, T. Tran, Y. Gal, P. H. Torr, and A. G. Baydin. Kl guided domain adaptation. *arXiv preprint arXiv:2106.07780*, 2021.

[127] A. Q. Nichol and P. Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021.

[128] P. Novosad, V. Fonov, and D. L. Collins. Unsupervised domain adaptation for the automated segmentation of neuroanatomy in MRI: a deep learning approach. *bioRxiv*, page 845537, 2019.

[129] T. Olsson, L. F. Barcellos, and L. Alfredsson. Interactions between genetic, lifestyle and environmental risk factors for multiple sclerosis. *Nature Reviews Neurology*, 13(1):25–36, 2017.

[130] M. Orbes-Arteaga, T. Varsavsky, C. H. Sudre, Z. Eaton-Rosen, L. J. Haddow, L. Sørensen, M. Nielsen, A. Pai, S. Ourselin, M. Modat, et al. Multi-domain adaptation in brain MRI through paired consistency and adversarial learning. In *Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data*, pages 54–62. Springer, 2019.

[131] Z. Pan and J. Lu. A bayes-based region-growing algorithm for medical image segmentation. *Computing in science & Engineering*, 9(4):32–38, 2007.

[132] J. L. Panman, Y. Y. To, E. L. van der Ende, J. M. Poos, L. C. Jiskoot, L. H. Meeter, E. G. Dopper, M. J. Bouts, M. J. Van Osch, S. A. Rombouts, et al. Bias introduced by multiple head coils in MRI research: an 8 channel and 32 channel coil comparison. *Frontiers in neuroscience*, page 729, 2019.

[133] S. Park and N. Kwak. Analysis on the dropout effect in convolutional neural networks. In *Asian conference on computer vision*, pages 189–204. Springer, 2016.

[134] D. D. Pham, G. Dovletov, and J. Pauli. Liver segmentation in CT with MRI data: zero-shot domain adaptation by contour extraction and shape priors. In *2020 IEEE 17th international symposium on biomedical imaging (ISBI)*, pages 1538–1542. IEEE, 2020.

[135] R. Pomponio, G. Erus, M. Habes, J. Doshi, D. Srinivasan, E. Mamourian, V. Bashyam, I. M. Nasrallah, T. D. Satterthwaite, Y. Fan, L. J. Launer, C. L. Masters, P. Maruff, C. Zhuo, H. Völzke, S. C. Johnson, J. Fripp, N. Koutsouleris, D. H. Wolf, R. Gur, R. Gur, J. Morris, M. S. Albert, H. J. Grabe, S. M. Resnick, R. N. Bryan, D. A. Wolk, R. T. Shinohara, H. Shou, and C. Davatzikos. Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan. *NeuroImage*, 208:116450, 2020.

[136] K. Preechakul, N. Chatthee, S. Wizadwongsa, and S. Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10619–10629, 2022.

[137] J. B. Putnam. Management of malignant pleural effusion: Sclerosis or chronic tube drainage. In *Difficult Decisions in Thoracic Surgery*, pages 414–423. Springer, 2007.

[138] K. Qin, K. Xu, F. Liu, and D. Li. Image segmentation based on histogram analysis utilizing the cloud model. *Computers & Mathematics with Applications*, 62(7):2824–2833, 2011.

[139] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[140] P. Rajpurkar, J. Irvin, A. Bagul, D. Ding, T. Duan, H. Mehta, B. Yang, K. Zhu, D. Laird, R. L. Ball, et al. Mura: Large dataset for abnormality detection in musculoskeletal radiographs. *arXiv preprint arXiv:1712.06957*, 2017.

[141] A. D. Rasamoelina, F. Adjailia, and P. Sinčák. A review of activation function for artificial neural network. In *2020 IEEE 18th World Symposium on Applied Machine Intelligence and Informatics (SAMI)*, pages 281–286, 2020.

[142] S. Raschka. Logistic regression and multi-class classification. https://sebastianraschka.com/pdf/lecture-notes/stat453ss21/L08_logistic__slides.pdf, 2020. Accessed on 20.07.2022.

[143] K. Raza and N. K. Singh. A tour of unsupervised deep learning for medical image analysis. *Current Medical Imaging*, 17(9):1059–1077, 2021.

[144] J. Ren, I. Hacihaliloglu, E. A. Singer, D. J. Foran, and X. Qi. Unsupervised domain adaptation for classification of histopathology whole-slide images. *Frontiers in bioengineering and biotechnology*, 7:102, 2019.

[145] D. Rezende and S. Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015.

[146] H. Robbins and S. Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[147] 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*, pages 234–241. Springer, 2015.

[148] F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

[149] F. Rosenblatt. Principles of neurodynamics. perceptrons and the theory of brain mechanisms. Technical report, Cornell Aeronautical Lab Inc Buffalo NY, 1961.

[150] L. Ruff, R. A. Vandermeulen, N. Görnitz, A. Binder, E. Müller, K.-R. Müller, and M. Kloft. Deep semi-supervised anomaly detection. *arXiv preprint arXiv:1906.02694*, 2019.

[151] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[152] P. Russo. *Handbook of X-ray imaging: physics and technology*. CRC press, 2017.

[153] A. D. Sadovnick, G. C. Ebers, R. W. Wilson, and D. W. Paty. Life expectancy in patients attending multiple sclerosis clinics. 42(5):991–991, 1992.

[154] C. Saharia, W. Chan, H. Chang, C. A. Lee, J. Ho, T. Salimans, D. J. Fleet, and M. Norouzi. Palette: Image-to-image diffusion models. *arXiv preprint arXiv:2111.05826*, 2021.

[155] M. Salem, S. Taheri, and J. S. Yuan. Anomaly generation using generative adversarial networks in host-based intrusion detection. In *2018 9th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, pages 683–687. IEEE, 2018.

[156] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training GANs. *Advances in neural information processing systems*, 29, 2016.

[157] N. Salman, B. Ghafour, and G. Hadi. Medical image segmentation based on edge detection techniques. *Advances in Image and Video Processing*, 3(2):1–9, 2015.

[158] S. Sankaranarayanan, Y. Balaji, C. D. Castillo, and R. Chellappa. Generate to adapt: Aligning domains using generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8503–8512, 2018.

[159] C. M. Scannell, A. Chiribiri, and M. Veta. Domain-adversarial learning for multi-centre, multi-vendor, and multi-disease cardiac mr image segmentation. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, pages 228–237. Springer, 2020.

[160] H. H. Schild. *MRI made easy (... well almost)*. Berlex Laboratories, 1992.

[161] 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*, pages 146–157. Springer, 2017.

[162] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[163] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[164] S. D. Serai, M.-L. Ho, M. Artunduaga, S. S. Chan, and G. B. Chavhan. Components of a magnetic resonance imaging system and their relationship to safety and image quality. *Pediatric radiology*, 51(5):716–723, 2021.

[165] S. Sharma, S. Sharma, and A. Athaiya. Activation functions in neural networks. *towards data science*, 6(12):310–316, 2017.

[166] J.-H. Shu, F.-D. Nian, M.-H. Yu, and X. Li. An improved mask r-cnn model for multiorgan segmentation. *Mathematical Problems in Engineering*, 2020, 2020.

[167] N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni. U-net and its variants for medical image segmentation: A review of theory and applications. *IEEE Access*, 2021.

[168] M. M. R. Siddiquee, Z. Zhou, N. Tajbakhsh, R. Feng, M. B. Gotway, Y. Bengio, and J. Liang. Learning fixed points in generative adversarial networks: From image-to-image translation to disease detection and localization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 191–200, 2019.

[169] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[170] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[171] K. Singh and S. Upadhyaya. Outlier detection: applications and techniques. *International Journal of Computer Science Issues (IJCSI)*, 9(1):307, 2012.

[172] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.

[173] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[174] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[175] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[176] R. Strudel, R. Garcia, I. Laptev, and C. Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7262–7272, 2021.

[177] A. J. Thompson, B. L. Banwell, F. Barkhof, W. M. Carroll, T. Coetzee, G. Comi, J. Correale, F. Fazekas, M. Filippi, M. S. Freedman, et al. Diagnosis of multiple sclerosis: 2017 revisions of the mcdonald criteria. *The Lancet Neurology*, 17(2):162–173, 2018.

[178] B. D. Trapp and K.-A. Nave. Multiple sclerosis: an immune or neurodegenerative disorder? *Annu. Rev. Neurosci.*, 31:247–269, 2008.

[179] A. M. Turing and J. Haugeland. Computing machinery and intelligence. *The Turing Test: Verbal Behavior as the Hallmark of Intelligence*, pages 29–56, 1950.

[180] T. Tykocki and M. Eltayeb. Ten-year survival in glioblastoma. a systematic review. *Journal of Clinical Neuroscience*, 54:7–13, 2018.

[181] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017.

[182] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

[183] N. Upadhyay and A. Waldman. Conventional MRI evaluation of gliomas. *The British journal of radiology*, 84(special_issue_2):S107–S111, 2011.

[184] S. Valverde, M. Salem, M. Cabezas, D. Pareto, J. C. Vilanova, L. Ramió-Torrentà, À. Rovira, J. Salvi, A. Oliver, and X. Lladó. One-shot domain adaptation in multiple sclerosis lesion segmentation using convolutional neural networks. *NeuroImage: Clinical*, 21:101638, 2019.

[185] D. C. Van Essen, K. Ugurbil, E. Auerbach, D. Barch, T. E. Behrens, R. Bucholz, A. Chang, L. Chen, M. Corbetta, S. W. Curtiss, et al. The Human Connectome Project: a data acquisition perspective. *Neuroimage*, 62(4):2222–2231, 2012.

[186] A. Van Oord, N. Kalchbrenner, and K. Kavukcuoglu. Pixel recurrent neural networks. In *International conference on machine learning*, pages 1747–1756. PMLR, 2016.

[187] A. Vasilev, V. Golkov, M. Meissner, I. Lipp, E. Sgarlata, V. Tomassini, D. K. Jones, and D. Cremers. q-space novelty detection with variational autoencoders. In *Computational Diffusion MRI*, pages 113–124. Springer, 2020.

[188] S. Vijayalakshmi et al. Image-guided surgery through internet of things. In *Internet of Things in Biomedical Engineering*, pages 75–116. Elsevier, 2019.

[189] J. Wang and X. Liu. Medical image recognition and segmentation of pathological slices of gastric cancer based on deeplab v3+ neural network. *Computer Methods and Programs in Biomedicine*, 207:106210, 2021.

[190] M. Wang and W. Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018.

[191] M. Weller, M. Van Den Bent, K. Hopkins, J. C. Tonn, R. Stupp, A. Falini, E. Cohen-Jonathan-Moyal, D. Frappaz, R. Henriksson, C. Balana, et al. Eano guideline for the diagnosis and treatment of anaplastic gliomas and glioblastoma. *The lancet oncology*, 15(9):e395–e403, 2014.

[192] M. Weller, W. Wick, K. Aldape, M. Brada, M. Berger, S. M. Pfister, R. Nishikawa, M. Rosenthal, P. Y. Wen, R. Stupp, et al. Glioma. *Nature reviews Disease primers*, 1(1):1–18, 2015.

[193] P. H. Winston. *Artificial intelligence*. Addison-Wesley Longman Publishing Co., Inc., 1992.

[194] J. Wolleb, F. Bieder, R. Sandkühler, and P. C. Cattin. Diffusion models for medical anomaly detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 35–45. Springer, 2022.

[195] J. Wolleb, R. Sandkühler, F. Bieder, M. Barakovic, N. Hadjikhani, A. Papadopoulou, Ö. Yaldizli, J. Kuhle, C. Granziera, and P. C. Cattin. Learn to ignore: domain adaptation for multi-site MRI analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 725–735. Springer, 2022.

[196] J. Wolleb, R. Sandkühler, F. Bieder, and P. C. Cattin. The swiss army knife for image-to-image translation: Multi-task diffusion models. *arXiv preprint arXiv:2204.02641*, 2022.

[197] J. Wolleb, R. Sandkühler, F. Bieder, P. Valmaggia, and P. C. Cattin. Diffusion models for implicit image segmentation ensembles. In *International Conference on Medical Imaging with Deep Learning*, pages 1336–1348. PMLR, 2022.

[198] J. Wolleb, R. Sandkühler, and P. C. Cattin. Descargan: Disease-specific anomaly detection with weak supervision. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 14–24. Springer, 2020.

[199] J. Wrobel, M. Martin, R. Bakshi, P. Calabresi, M. Elliot, D. Roalf, R. Gur, R. Gur, R. Henry, G. Nair, J. Oh, N. Papinutto, D. Pelletier, D. Reich, W. Rooney, T. Satterthwaite, W. Stern, K. Prabhakaran, N. Sicotte, R. Shinohara, and J. Goldsmith. Intensity warping for multisite mri harmonization. *NeuroImage*, 223:117242, 2020.

[200] Y. Wu and K. He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[201] X. Xia, X. Pan, N. Li, X. He, L. Ma, X. Zhang, and N. Ding. GAN-based anomaly detection: A review. *Neurocomputing*, 2022.

[202] W.-T. Xiao, L.-J. Chang, and W.-M. Liu. Semantic segmentation of colorectal polyps with deeplab and lstm networks. In *2018 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)*, pages 1–2. IEEE, 2018.

[203] H. Yan, Y. Ding, P. Li, Q. Wang, Y. Xu, and W. Zuo. Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2272–2281, 2017.

[204] J. Yang, W. An, S. Wang, X. Zhu, C. Yan, and J. Huang. Label-driven reconstruction for domain adaptation in semantic segmentation. In *European conference on computer vision*, pages 480–498. Springer, 2020.

[205] J. Yang, R. Xu, Z. Qi, and Y. Shi. Visual anomaly detection for images: A survey. *arXiv preprint arXiv:2109.13157*, 2021.

[206] L. Zhang, X. Wang, D. Yang, T. Sanford, S. Harmon, B. Turkbey, B. J. Wood, H. Roth, A. Myronenko, D. Xu, et al. Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. *IEEE transactions on medical imaging*, 39(7):2531–2540, 2020.

[207] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[208] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[209] 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.

[210] F. E. Zink. X-ray tubes. *Radiographics*, 17(5):1259–1268, 1997.

Curriculum Vitae

Julia Wolleb

Education

2019–2022	**Ph.D. in Biomedical Engineering** Department of Biomedical Engineering, University of Basel Topic: *Automatic Detection of Pathological Regions in Medical Images* Supervisors: Prof. Dr. Philippe C. Cattin, Prof. Dr. Cristina Granziera
2016–2018	**Master of Science in Mathematics**, University of Basel Master's thesis: *Modelling and Simulation of the Dispersal of Aedes albopictus across Switzerland* at the Swiss Tropical and Public Health Institute Supervisors: Prof. Dr. Marcus Grote, PD Dr. Nakul Chitnis
2013–2016	**Bachelor of Science in Mathematics**, University of Basel

List of Publications

Peer-reviewed Publications

2022 J. Wolleb, R. Sandkühler, F. Bieder, M. Barakovic, N. Hadjikhani, A. Papadopoulou, Ö. Yaldizli, J. Kuhle, C. Granziera, and P. C. Cattin. Learn to ignore: domain adaptation for multi-site MRI analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 725–735. Springer, 2022

J. Wolleb, F. Bieder, R. Sandkühler, and P. C. Cattin. Diffusion models for medical anomaly detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 35–45. Springer, 2022

J. Wolleb, R. Sandkühler, F. Bieder, P. Valmaggia, and P. C. Cattin. Diffusion models for implicit image segmentation ensembles. In *International Conference on Medical Imaging with Deep Learning*, pages 1336–1348. PMLR, 2022

F. Bieder, J. Wolleb, R. Sandkühler, and P. C. Cattin. Position regression for unsupervised anomaly detection. In *International Conference on Medical Imaging with Deep Learning*, pages 160–172. PMLR, 2022

2020 J. Wolleb, R. Sandkühler, and P. C. Cattin. Descargan: Disease-specific anomaly detection with weak supervision. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 14–24. Springer, 2020

Technical Reports

2022 J. Wolleb, R. Sandkühler, F. Bieder, and P. C. Cattin. The swiss army knife for image-to-image translation: Multi-task diffusion models. *arXiv preprint arXiv:2204.02641*, 2022